UNIVERSIDAD DEL NORTE

Departamento de Ingeniería Mecánica

DISEÑO DE UNA TÉCNICA MULTIVARIADA DE PROCESAMIENTO DE DATOS PARA CARACTERIZACIÓN DE ESTRUCTURAS ESTADÍSTICAS SUBYACENTES APLICABLE A SISTEMAS BIOMÉDICOS E INDUSTRIALES

TESIS

Para la obtención del título de:

Doctor en Ingeniería Mecánica

IVÁN DARÍO PORTNOY DE LA OSSA

Barranquilla, Colombia

Abril 2020

UNIVERSIDAD DEL NORTE

Departamento de Ingeniería Mecánica

DISEÑO DE UNA TÉCNICA MULTIVARIADA DE PROCESAMIENTO DE DATOS PARA CARACTERIZACIÓN DE ESTRUCTURAS ESTADÍSTICAS SUBYACENTES APLICABLE A SISTEMAS BIOMÉDICOS E INDUSTRIALES

TESIS

Para la obtención del título de:

Doctor en Ingeniería Mecánica

IVÁN DARÍO PORTNOY DE LA OSSA

Barranquilla, Colombia

Abril 2020

AGRADECIMIENTOS

Este trabajo no podría haber sido realizado sin la indispensable guía de mis asesores. Las enriquecedoras discusiones intelectuales con el Prof. Marco Sanjuán, y la guía y asesoría incondicional del Prof. Eduardo Zurek constituyeron una contribución inconmensurable para esta investigación. Debo también agradecer al Dr. Homero San Juan por sus valiosos aportes a esta investigación. Agradezco también a Sebastian Racedo por su colaboración y apoyo. Finalmente, a mi familia, mi novia y mis amigos por el apoyo emocional.

FINANCIAMIENTO

El estudio de doctorado, en cuyo marco se desarrolla esta investigación, fue financiado por COLCIENCIAS y la Gobernación del Atlántico (Colombia) a través del crédito condonable No. 673 (2014), "*Formación de Capital Humano de Alto Nivel para el Departamento del Atlántico*".

Además, esta investigación fue parcialmente financiada por la subvención de COLCIENCIAS No. 1215-5693-4635, contrato 0770-2013, así como también por el proyecto No. R01AI110385 del National Institute of Allergy and Infectious Diseases, National Institutes of Health (NIH), Bethesda (Estados Unidos).

El contenido de este documento es responsabilidad exclusiva del autor y sus directores de tesis y asesores, y no representa necesariamente las opiniones de ninguna institución o agencia patrocinadora. Los patrocinadores del estudio no tuvieron ninguna participación en el diseño del estudio, recolección de datos, análisis de datos, interpretación de datos o en la escritura del documento.

TABLA DE CONTENIDO

LISTADO DE TABLAS

LISTADO DE TABLAS

ABREVIACIONES

AIE: Average (variable) identification effectiveness

ADN: Ácido desoxirribonucleico

AMR: Average misdetection rate

BM: Bayesian multiplicative (Algorithm)

bp: Base pair

CLR: Centered log-ratio

ddNTP: Didesoxinuleótido

DESeq2: Differential expression of RNA-Seq data 2

dna: Differential network analysis

FAO: Fixed airway obstruction

FDA: Fisher discriminant analysis

ICA: Independent component analysis

iDINGO: Integrative differential network analysis in genomics

INDEED: Integrated differential expression and differential network analysis

LASSO: Least absolute shrinkage and selection operator

LEfSe: Linear discriminant analysis effect size

NAO: No airway obstruction

NCA: Non-parametric compositional-data assessment

NGS: Next generation sequencing

NOC: Normal Operation Condition

OTU: Operational taxonomic unit

PCA: Principal component analysis

PCR: Polymerase chain reaction

PLS: Partial least squares

PLSDA: Partial least squares differential analysis

RAO: Reversible airway obstruction

rRNA: Ribosomal ribonucleic acid

SMRT: Single molecule real time (sequencing)

SPLSDA: Sparse PLSDA

SPIEC-EASI: Sparse inverse covariance estimation for ecological association inference

SSE: South Sparkling lake epilimnion

TBE: Trout Bog lake epilimnion

TBH: Trout Bog lake hypolimnion

TIER: Type I error rate

RESUMEN

El análisis de la estructura de correlación entre las variables que componen un sistema (e.g., un proceso industrial, un organismo vivo, etc.) permite el discernimiento de las interacciones subyacentes de dicho sistema, así como también la caracterización de sus distintas configuraciones o estados. Esto permite crear herramientas para la detección de anomalías basándose en la detección de cambios en la estructura de correlación de estos sistemas (Jiang and Braatz, 2017). En el campo del control automático de procesos y la detección y diagnóstico de fallas se han desarrollado técnicas que monitorean la estructura de correlación, estudiando los cambios estructurales en ésta para llevar a cabo la detección y diagnóstico de condiciones atípicas de operación o transiciones entre distintas condiciones de operación. Sin embargo, la incorporación del análisis de esta estructura de correlación aún es incipiente, por lo que hay un creciente interés en su uso en este campo (Liu, Zhong and Ma, 2013; Mayer-Schönberger and Cukier, 2013; Rato and Reis, 2014; Severson, Chaiwatanodom and Braatz, 2016; Reis and Gins, 2017).

En el contexto de los sistemas biológicos, la estructura de redes de correlación (entre otras métricas de asociación), tal como niveles de expresión de genes o poblaciones microbianas, por ejemplo, pueden proporcionar una idea sobre las interacciones biológicas subyacentes que tienen lugar dentro de estos sistemas. Por lo tanto, el desarrollo de algoritmos para la (re)construcción de redes biológicas se ha convertido en un tema de investigación de relevancia (Gill, Datta and Datta, 2010; Mangan *et al.*, 2016; Muetze *et al.*, 2016; Montagud *et al.*, 2019; Sonawane *et al.*, 2019; Treur, 2019).

Con respecto a los procesos industriales, aquellos que posean un gran número de variables que se correlacionen entre sí y que además estén altamente instrumentados (i.e., que posean sensores/transmisores y sistemas de adquisición y almacenamiento de datos), son adecuados para la implementación y validación de técnicas como las diseñadas en este trabajo. Tal es el caso de procesos como los termoquímicos, procesos

de manufactura altamente instrumentados, distritos de frío, sistemas de transporte de hidrocarburos, etc. La idoneidad de este tipo de procesos para la implementación y validación de este tipo de técnicas se debe a que: i) Se tiene disponibilidad de datos históricos, así como también de datos en tiempo real (on-line), permitiendo el entrenamiento y ejecución de las técnicas, y ii) a partir de la estructura de correlaciones de las variables se extrae información valiosa que permite caracterizar los modos de operación y de fallas (comportamientos anómalos) de estos procesos.

La correlación de Pearson (Pearson, 1897) es la métrica más popular para medir la asociación entre un par de variables. Esta métrica, sin embrago, presenta algunas limitaciones y no puede ser aplicada de manera directa para analizar cualquier tipo de datos como los de naturaleza composicional (Pearson, 1897), ya que los resultados son propensos a exhibir correlaciones espurias. Lo anterior se debe a que los datos de entrada para el estadístico están restringidos a sumar una cantidad constante, por lo que no es posible que una variable cambie sin alterar (artificialmente) al menos otra variable. Tal es el caso para los datos composicionales de expresión de genes, datos de transcriptómica, así como también el de las matrices de conteos de unidades operacionales taxonómicas (OTUs, por sus siglas en inglés) (Van Dongen, Abreu-Goodger and Enright, 2008; Erb and Notredame, 2016). Por consiguiente, métodos estadísticos estándares no deben ser usados para analizar datos composicionales sin consideraciones especiales.

Aunque en la literatura existe un amplio espectro de métodos para la construcción de redes bilógicas, la mayoría de ellos son incapaces de manejar datos composicionales correctamente (Juric *et al.*, 2007). Sin embargo, se han desarrollado técnicas recientes para analizar este tipo de datos, tales como el algoritmo denominado SPIEC-EASI (sparse inverse covariance estimation for ecological association inference), el cual reconstruye la red de correlaciones basado en el algoritmo de vecindad dispersa (sparse neighborhood algorithm) y en la covarianza inversa (Kurtz *et al.*, 2015).

En este trabajo se desarrolla y ejemplifica la técnica *non-parametric compositional-data assessment* (NCA), un nuevo método para la discriminación cuantitativa de estructuras de correlación y el análisis diferencial en datos provenientes de sistemas biológicos. NCA puede procesar, pero no se limita a, datos de naturaleza composicional, permitiendo determinar aquellas variables que contribuyen en mayor medida a las desviaciones en las estructuras de correlación. Se introducen, con NCA, un estadístico para cuantificar esas desviaciones entre dos conjuntos de datos, junto con un procedimiento basado en bootstrapping para evaluar significancia estadística en las diferencias encontradas.

Experimentos de simulación, con datos sintéticos, muestran que NCA supera a otros métodos disponibles en la literatura en términos del error tipo I y la potencia estadística, especialmente cuando aumenta el número de variables. Por esto, NCA es adecuado para el análisis de datos de gran dimensión provenientes de experimentos de secuenciación, aunque su extensión a otros tipos de datos es sencilla. Finalmente, se ilustra la aplicación de NCA con tres casos estudio usando datos provenientes de experimentos de secuenciación.

1. INTRODUCCIÓN

En este capítulo se discute, en primer lugar, la relevancia y los retos de incluir el análisis de la estructura de correlación de las variables que componen un sistema, sea un proceso termoquímico, mecánico, eléctrico, etc., o un sistema biológico. Luego, se evidencian las limitaciones de las técnicas de análisis de datos para datos composicionales, enfocándose particularmente en datos biológicos resultantes de experimentos de secuenciación. De esta manera, se plantea el problema de investigación, resaltándose posteriormente la relevancia de éste. Se plantean finalmente los objetivos, general y específicos, cuyo cumplimiento está enfocado a la resolución del problema de investigación planteado.

1.1. Planteamiento del Problema

Un sistema es un grupo de entidades que interactúan o se interrelacionan para formar un todo. Un sistema es delineado por sus fronteras espaciales y temporales, es rodeado e influenciado por su ambiente, es declarado en su funcionamiento, y es descrito por su propósito y estructura. El análisis de la estructura de correlación entre las variables que componen un sistema es fundamental para un entendimiento exhaustivo de las distintas configuraciones estructurales de dicho sistema, así como también para el desarrollo de herramientas de detección y diagnóstico de configuraciones anómalas en éste (Jiang and Braatz, 2017). Sin embargo, la estructura de correlación todavía no ha sido explorada extensivamente para desarrollar este tipo de aplicaciones, por lo que hay aún un creciente interés en su uso en este campo (Liu, Zhong and Ma, 2013; Severson, Chaiwatanodom and Braatz, 2016; Montagud *et al.*, 2019; Sonawane *et al.*, 2019; Treur, 2019). Algunas técnicas, desarrolladas recientemente en el campo del control automático y monitoreo de procesos, buscan la identificación de cambios o distorsiones en la estructura de correlación de las variables para el discernimiento de cambios en las condiciones de operación de sistemas, así como también para la

detección y diagnóstico de condiciones atípicas de operación, denominadas fallas (Rato and Reis, 2014; Reis and Gins, 2017).

Si se hace la analogía de una célula con un proceso industrial, el material genético de las células es una fuente de información cuyo procesamiento, uso, mantenimiento y copia son procesos fundamentales que pueden ser monitoreados, estudiados y modelados, así como también se puede hallar los subsistemas principalmente afectados durante una condición atípica (e.g., infección viral) y cómo cambian éstos a lo largo del desarrollo de la enfermedad. Esta analogía ofrece un parangón entre los sistemas biológicos y los procesos industriales de los cuales se tengan registros históricos de datos y disponibilidad de almacenamiento de datos de operación en línea. Tal es el caso de los procesos como los termoquímicos, procesos de manufactura, distritos de frío, sistemas de transporte de hidrocarburos, etc. (Chiang, Russell and Braatz, 2000). Estos procesos, además de estar altamente instrumentados con sensores/transmisores que recopilan y envían sus señales a sistemas de adquisición de datos, exhiben, por naturaleza, correlaciones entre sus múltiples variables, ofreciendo un insumo valioso para extraer información de su estructura de correlaciones que permite caracterizar los modos de operación (normales o anómalos) de estos procesos éstos e implementar técnicas similares a aquellas propuestas en este trabajo.

Con respecto a los sistemas biológicos, la introducción de plataformas masivas de secuenciación de última generación (NGS) ha permitido secuenciar simultáneamente cientos de miles de fragmentos de ADN, cambiando el panorama de los estudios de genética debido a la gran cantidad de datos disponibles para ser analizados (Weng, Zhang and Zhang, 2004; Kim, Golub and Park, 2006; Carin *et al.*, 2012; Saha *et al.*, 2013). Con el rápido desarrollo de tecnologías de secuenciación de alto rendimiento hay una acumulación vertiginosa de datos recolectados de procesos biológicos, incluyendo datos de genómica, transcriptómica, proteómica, metabolómica e interactómica (OMICs) (Zhang *et al.*, 2014). La era de las OMICs ha traído grandes avances en el desciframiento de algunas enfermedades humanas. Sin embargo, la gran

complejidad de los mecanismos biológicos subyacentes a dichas enfermedades continúa presentando retos a los investigadores.

El estado de salud de un individuo puede alterar el perfil de expresión de genes de una célula, tal como ocurre cuando se está desarrollando una infección viral o se tiene un tumor canceroso (Carin *et al.*, 2012; Lou and Obradovic, 2012), por lo que resulta pertinente caracterizar y estudiar dichos perfiles de expresión de genes, así como también caracterizar su evolución temporal en las distintas etapas del cuadro infeccioso. Estos datos deben ser analizados de manera conjunta y capturar las correlaciones entre ellos, con lo cual se comienza a prescindir de la errónea concepción de que cualquier enfermad humana puede ser atribuida a un solo marcador molecular específico (Munos, 2009; Paul *et al.*, 2010).

Aunque los métodos de estadística multivariada para la reducción dimensional, así como los métodos de reconocimiento de patrones, son concebidos desde áreas como la ingeniería, su versatilidad hace que sean apropiados para las aplicaciones no solamente en las ciencias de la salud sino también en otras áreas científico-técnicas, aportando tecnologías de soporte al desarrollo tecnológico del sector industrial y de la salud. Uno de los retos más relevantes en el análisis de expresión de genes consiste en la identificación de aquellos genes que exhiben cambios significativos en su nivel de expresión cuando se presentan condiciones atípicas, tales como patologías o condiciones clínicas anormales. Este tipo de análisis es conocido como *análisis diferencial*. Para llevar a cabo este tipo de análisis se hace necesaria la implementación de algoritmos de reducción dimensional, que buscan disminuir la complejidad y demanda computacional requerida para el procesamiento de datos. Además, la estructura de correlación entre la gran cantidad de variables disponibles debe ser integrada a la base de conocimiento de los algoritmos de análisis o a la mecánica misma de procesamiento, teniendo además consideraciones concernientes a la naturaleza composicional de los datos.

Como se evidencia en la revisión del estado del arte, la mayoría de las técnicas que se usan para análisis de correlaciones y análisis diferencial, aplicadas a datos provenientes de secuenciación genética, ignoran la naturaleza composicional de este tipo de datos. Ya en 1897 el matemático Karl Pearson advirtió que la geometría Euclidiana no es apropiada para el análisis de datos composicionales, en particular cuando se trata de análisis de correlación, ya que surgen correlaciones espurias entre proporciones de medidas absolutas (Pearson, 1897). Lo anterior se debe a que los datos están restringidos a sumar una cantidad constante, por lo que no es posible que una variable cambie sin alterar (artificialmente) al menos otra variable. Se requiere entonces un enfoque diferente, que sea sensible a la naturaleza composicional de estos datos. Se evidencia también una carencia de métodos que comparen cuantitativamente la estructura de correlación de grupos disímiles cuando los datos son de esta naturaleza.

Debido a lo anterior, se demuestra la pertinencia de llevar a cabo una investigación para desarrollar técnicas multivariadas de procesamiento de datos de expresión genética en células bajo infección del viral para la caracterización de su estructura estadística subyacente, desarrollándose además indicadores para técnicas de análisis de datos de expresión de genes. Este tipo de técnicas podrían también ser aplicadas en una diversa gama de campos científicos y técnicos, tales como el procesamiento de señales de proceso, análisis de big data, geología, química, ciencias forenses, sociología, economía, etc., ya que existen datos de naturaleza composicional en estos campos.

1.2. Objetivos de la Investigación

En esta sección se define el objetivo general, así como también los objetivos específicos de la investigación.

1.2.1. Objetivo General

Desarrollar una técnica para el análisis de estructuras de correlación y análisis diferencial para datos composicionales con el fin de estimar la estructura subyacente de la información obtenida, validada en el contexto de datos de expresión genética en pacientes con patologías de interés.

1.2.2. Objetivos Específicos

- Desarrollar una técnica de reducción dimensional de muestras provenientes de experimentos de secuenciación que considere la naturaleza composicional de los datos.
- Desarrollar técnicas de análisis diferencial y de discriminación/diferenciación cuantitativa de matrices de correlación que consideren la naturaleza composicional de los datos de expresión genética.
- Desarrollar una metodología de integración de las técnicas diseñadas que permita un análisis robusto de datos de expresión genética.
- Desarrollar indicadores para la evaluación del desempeño de la metodología propuesta, permitiendo además la comparación con técnicas del estado del arte.

2. REVISIÓN DEL ESTADO DEL ARTE

Se lleva a cabo en este capítulo una revisión del estado del arte concerniente a los siguientes temas: Técnicas de secuenciación, técnicas de reducción dimensional, técnicas de análisis diferencial, y técnicas de evaluación de similitud de estructuras de correlación.

2.1. Técnicas de Secuenciación

El primer protocolo de secuenciación de ADN (ácido desoxirribonucleico) fue desarrollado por Allan Maxam y Walter Gilbert (Maxam and Gilbert, 1977) en la Universidad de Harvard en 1977. Maxam y Gilbert denominaron *digestión química* a su protocolo, el cual consistía en la marcación de los extremos (5' o 3') de una o ambas hebras de ADN con 32P, un isotopo radiactivo del fosforo, y propiciar cuatro distintas reacciones que tienen lugar en distintos recipientes en los que se ha dividido la muestra de manera equitativa. En cada una de estas reacciones, se hacen cortes en las hebras en nucleótidos específicos marcados, obteniéndose al final hebras con distintas longitudes que demarcan los lugares ocupados por el nucleótido objetivo (para cada reacción). La digestión química permitía la secuenciación de cadenas de aproximadamente 100 pares de bases (bp).

Sanger et al. (Sanger, Nicklen and Coulson, 1977) propusieron, también en 1977, el método de secuenciación *Sanger*, el cual se basa en cuatro reacciones de síntesis de ADN que tienen lugar en cuatro tubos distintos. Para cada reacción se usa la ADN-polimerasa (encima que propicia la síntesis) nucleótidos (adenina: A; Guanina: G; Timina: T; Citocina: C) y didesoxinucleótidos (ddNTP) específicos para cada reacción, los cuales están marcados con 32P y además carecen de un grupo hidroxilo en su extremo 3'. Al agregar la polimerasa en cada tubo, que ya contiene los ddNTPs específicos y los demás nucleótidos, se empezarán a acoplar nucleótidos hasta que la síntesis es interrumpida al acoplar el ddNTP. Este proceso ocurre de manera aleatoria,

produciendo, en cada tubo, cadenas de todas las longitudes posibles que terminan en el ddNTP específico. Posteriormente, los fragmentos obtenidos se separan en un gel de poliacrilamida en cuatro carriles distintos usando electroforesis. El método Sanger original permitía la secuenciación de cadenas de hasta 300 bp.

En 1987, Hood et al. (Hood, Hunkapiller and Smith, 1987) desarrollaron el primer equipo comercial, basado en la tecnología Sanger (con algunas modificaciones), que podía ejecutar la secuenciación de 200bp por hora (por muestra) de manera automática, el Applied Biosystems 370A. Dentro de la siguiente generación de máquinas automáticas de secuenciación, se debe resaltar el ABI PRISM 3700 (Venter *et al.*, 1998), que podía secuenciar hasta 550 bp por cada reacción de secuenciación cada 2.5 horas. El ABI PRISM 3700 fue el secuenciador usado para el proyecto genoma humano (Collins and Mansoura, 2001), el cual tuvo un costo de unos €2 047 millones y tomó más de 10 años.

El siguiente salto tecnológico surge con las tecnologías de secuenciación de próxima generación (o NGS, por sus siglas en inglés), las cuales se basan en la lectura de múltiples secuencias cortas (~100bp) en paralelo, arrojando millones de lecturas al mismo tiempo, y a un costo mucho menor (Shendure and Ji, 2008; Metzker, 2010; Buermans and Den Dunnen, 2014). Para ejemplificar la reducción de costos, considérese lo siguiente: según el NIH (National Institutes of Health), en 1990 el costo de secuenciación era de aproximadamente $10/nucleótido, mientras que en el 2005 era de tan solo $0.01/nucleótido (tomado de https://www.genome.gov/27541954/dna-sequencing-costs-data/). La evolución temporal de los costos de secuenciación, en $/nucleótido, se muestra en la gráfica elaborada por el NIH (ver Figura 1) para el periodo comprendido entre 2001 y 2017.

Las tecnologías de NGS ofrecen una amplia variedad de métodos y configuraciones, pero algunos pasos previos a la lectura de las secuencias de ADN son comunes a todas, dentro de los cuales podemos encontrar: la disolución, la fragmentaciones de largas

cadenas en hebras de menor longitud, la adición de adaptadores, y la amplificación a través de la reacción en cadena de la polimerasa (o PCR, por sus siglas en inglés), la cual fue propuesta por Mullis et al. (Mullis *et al.*, 1986), y que constituye un hito tecnológico fundamental para la existencia de las técnicas NGS.

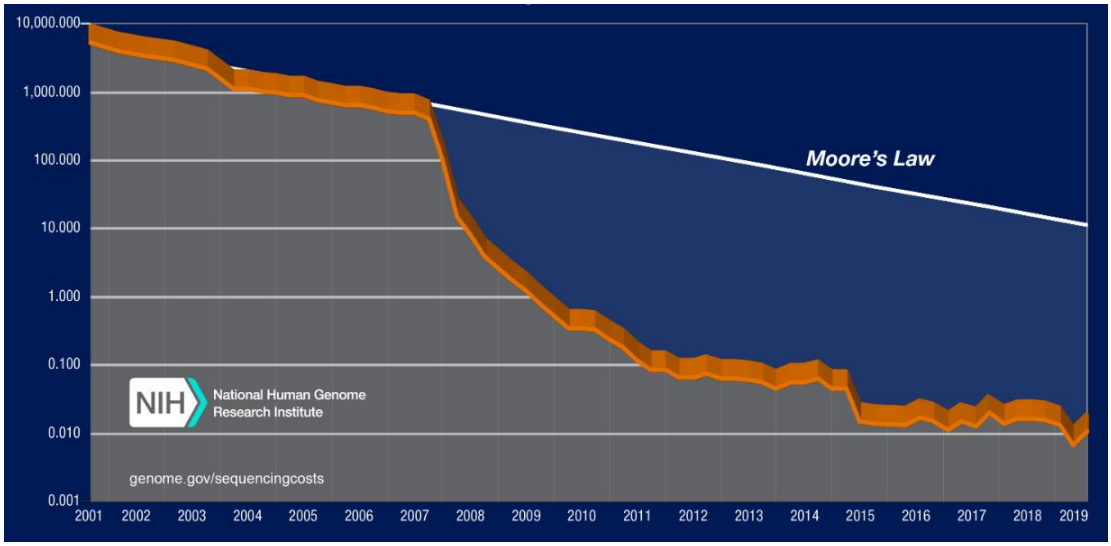

Figura 1 - Evolución Temporal de los costos (por Mega-Base) de secuenciación de ADN (Tomado de https://www.genome.gov/27541954/dna-sequencing-costs-data/).

Las principales diferencias entre los secuenciadores NGS se encuentran en la técnica empleada para la preparación del ADN (librerías) y el protocolo de secuenciación empleado (Rothberg *et al.*, 2011; Liu *et al.*, 2012; Goodwin, McPherson and McCombie, 2016). De acuerdo con la técnica de preparación, las tecnologías NGS se clasifican en: PCR en Emulsión, PCR en puente, ligación, secuenciación por semiconductores, y secuenciación por reversión del terminador.

Más recientemente, se ha desarrollado una tecnología que también pertenece a las NGS, denominada secuenciación de molécula simple en tiempo real (SMRT, por sus siglas en inglés). Esta tecnología usa adaptadores circulares y lleva a cabo la secuenciación a

través de un tipo especial de chips en los que se detecta el cambio en la actividad de la ADN-polimerasa, o a través de un cambio de voltaje debido al paso de la hebra de ADN por un poro, traduciendo posteriormente estos cambios en la secuencia que se desea leer (Roberts, Carneiro and Schatz, 2013, 2017; Ardui *et al.*, 2018)

2.2. Técnicas de Reducción Dimensional

Debido a que, como ya se ha dicho, los datos de expresión de genes son generalmente de gran dimensión (Kim, Golub and Park, 2006; Saha *et al.*, 2013) y el número de biomarcadores utilizados como características es típicamente mucho mayor que el número de sujetos sobre los que se realiza el estudio, deben usarse técnicas que permitan analizar gran cantidad de datos o en su defecto que permitan reducir la dimensión de dichos datos para su posterior análisis.

Se describen brevemente a continuación las principales técnicas de reducción dimensional encontradas en la revisión de la literatura, las cuales son usadas, entre otras aplicaciones, para detección temprana de fallas en entornos industriales, y en este caso, para analizar datos de expresión de genes (Zadeh, 1999; Shan and Deng, 2009).

2.2.1. Principal Component Analysis (PCA)

PCA es una técnica de reducción dimensional que captura óptimamente la variabilidad de los datos en un espacio de pequeña dimensión que es ampliamente usada en el monitoreo de procesos (MacGregor and Kourti, 1995; Choi and Lee, 2004; Amanian *et al.*, 2007) así como también en el análisis de datos de expresión de genes (Josserand, 2008; Shan and Deng, 2009). PCA permite además el conocimiento de las variables principalmente responsables de la estructura estadística subyacente a los datos (Jeng, 2010).

2.2.2. Independent Component Analysis (ICA)

El análisis de componentes independientes (ICA), propuesto por Herault y Jutten en 1968 (Herault and Jutten, 1986), consiste en buscar una transformación lineal que minimice la dependencia estadística entre los componentes de un conjunto de datos.

2.2.3. Fisher Discriminant Analysis (FDA)

FDA es también una técnica de reducción dimensional que es óptima para maximizar la separación entre distintas clases (patrones o estructuras). FDA tiene en cuenta la información entre las distintas clases (Xu, Yang and Yang, 2004).

2.2.4. Partial Least Squares (PLS)

Es, al igual que PCA, FDA e ICA, una técnica de reducción de dimensión, la cual busca maximizar la covarianza entre un bloque predictor y un bloque predicho para cada componente (Chiang, Russell and Braatz, 2000; Russell, Chiang and Braatz, 2012).

2.3. Técnicas de Análisis Diferencial

El análisis diferencial es un abordaje ampliamente usado en biología y ciencias médicas para la identificación de biomarcadores a través de la comparación de dos (o más) conjuntos de datos obtenidos bajo condiciones biológicas (e.g., patológicas) diferentes (Anders and Huber, 2010; Langmead, Hansen and Leek, 2010; Zhao and Qin, 2013).

A pesar de su utilidad, los métodos para el análisis diferencial basados en estadística multivariada todavía presentan algunas limitaciones. Por ejemplo, algunas técnicas se basan en métodos de estimación de redes (networks) singulares (Meinshausen, Bühlmann and others, 2006; Herd *et al.*, 2014; Reineberg *et al.*, 2018) o métodos de estimación conjunta (joint estimation) (Qiu *et al.*, 2016; Saegusa and Shojaie, 2016;

Wu *et al.*, 2019). Este tipo de técnicas no proporcionan estimaciones de incertidumbre (e.g., valores p).

Tabla 1 - Resumen de técnicas más recientes para el análisis diferencial de datos de secuenciación.

Técnicas de Análisis Diferencial para Datos provenientes de Experimentos de Secuenciación				
Año	**Técnica**	**Autor(es)**	**Provee Prueba de Hipótesis para Similitud de Estructuras de Correlación**	**Tratamiento para Datos Composicionales**
2010	SPLSDA	Chung & Keles	NO	NO
2010	Dna	Gill et al.	SI	NO
2011	LEfSe	Segata et al.	NO	NO
2014	DESeq2	Love et al.	NO	NO
2016	INDEED	Zuo et al.	NO	NO
2018	Estimación singular de redes	Reineberg et al.	NO	NO
2018	iDINGO	Class et al.	NO	NO
2019	Estimación conjunta de redes	Wu et al.	NO	NO

Por otro lado Love et al. (Love, Huber and Anders, 2014) propusieron el algoritmo DESeq2 (differential expression of RNA-Seq data 2) para el análisis de datos de expresión de genes. DESeq2 está diseñado para mejorar la clasificación y visualización de los genes marcadores, basándose en la estimación del tamaño de los recuentos de los genes, así como también para realizar pruebas estadísticas en los niveles de expresión y definir umbrales basados en la importancia biológica.

Segata et al. (Segata *et al.*, 2011) propusieron el algoritmo LEfSe (linear discriminant analysis effect size), el cual resalta la consistencia y relevancia biológicas de los efectos, identificando las variables diferencialmente abundantes que también están presentes en las categorías biológicas significativas.

Basándose en el método de mínimos cuadrados parciales para análisis discriminante (PLSDA, por sus siglas en inglés), Chung y Keles (Chung and Keles, 2010) desarrollaron el método SPLSDA (sparse PLSDA), una variante de PLSDA para datos dispersos, el cual supera a PLSDA en términos de su desempeño para hacer clasificación. La Tabla 1 muestra un resumen de estas técnicas, junto con algunos atributos notables de las mismas.

Zuo et al. (Zuo *et al.*, 2016) desarrollaron INDEED (INtegrated DiffErential Expression and Differential network analysis), una técnica que integra el análisis diferencial de datos de expresión con análisis de redes (networks) para el descubrimiento de biomarcadores, la cual se basa en la reconstrucción de una matriz dispersa de correlaciones mediante regresiones LASSO (Least Absolute Shrinkage and Selection Operator).

Class et al. (Class *et al.*, 2018) desarrollaron la técnica iDINGO (integrative differential network analysis in genomics), para estimar dependencias específicas de grupo y hacer inferencias en las redes diferenciales, considerando la jerarquía biológica entre diferentes plataformas.

2.4. Técnicas de Evaluación de Similitud de Estructuras de Correlación

Si bien las técnicas antes mencionadas para análisis diferencial (ver sección 2.2) constituyen herramientas útiles para el hallazgo de variables o marcadores que han cambiado significativamente su interacción relativa dentro de una red, también es importante tener una prueba estadística para evaluar la diferencia entre la estructura de red global (correlación) de dos conjuntos de datos, de tal manera que la incertidumbre se tenga en cuenta en el proceso de toma de decisiones.

Kullback (Kullback, 1967) desarrolló el primer método para evaluar la igualdad de la estructura de correlación de los conjuntos de datos basándose en el supuesto de que los datos se ajustan a las distribuciones normales multivariadas. Posteriormente, Cole (Cole, 1968) presentó la prueba de razón de verosimilitud (o *likelihood ratio test*, en inglés), que debe cumplir con el supuesto de normalidad y es sensible a violaciones del mismo (Modarres and Jernigan, 1993). Jennrich (Jennrich, 1970) propuso una prueba de hipótesis para la igualdad de matrices de correlación basada en el estadístico χ^2, también bajo la misma suposición.

Larntz y Pearlman in 1985 (Larntz and Perlman, 1985), basados de nuevo en el supuesto de normalidad, propusieron otra prueba para la igualdad de matrices de correlación. Modarres y Jernigan (Modarres and Jernigan, 1993), bajo la suposición de que los vectores de correlación siguen una distribución normal, desarrollaron otra prueba basada en el estadístico Q_m. El mismo año, Krzanowski (Krzanowski, 1993) presentó un enfoque no paramétrico basado en permutaciones para probar la igualdad de dos matrices de correlación bajo el supuesto de que los dos conjuntos de datos que se comparan siguen la misma distribución, cualquiera que sea ésta.

Tabla 2 - Resumen de técnicas existentes para la evaluación de igualdad de estructuras de correlación.

Técnicas para Prueba de Hipótesis de Igualdad de Estructuras de Correlación				
Año	**Autor(es)**	**Supuestos Sobre Distribución de Datos**	**Tratamiento para Datos Composicionales**	**Exploración de Tasas de Error Tipo I y II para Regiones Amplias de Dimensionalidad**
1967	Kullback	Distribución Normal	NO	NO
1968	Cole	Distribución Normal	NO	NO
1970	Jennrich	Distribución Normal	NO	NO
1985	Larntz & Pearlman	Distribución Normal	NO	NO

1993	Modarres & Jernigan	Vector de correlaciones sigue distribución normal	NO	NO
1993	Krzanowski	Ambos conjuntos de datos siguen la misma distribución	NO	NO
2014	Gill et al.	Ninguno	NO	Parcialmente (Sólo para p=20 y 100 variables)
2015	Ren et al.	Ninguno	NO	NO
2016	Belilovsky et al.	Ninguno	NO	NO
2017	Städler & Mukherjee	Ninguno	NO	NO
2017	Janková & van de Geer	Ninguno	NO	NO
2017	Xia & Li	Ninguno	NO	NO

Gill et al. (Gill, Datta and Datta, 2014) desarrollaron métodos, basados en re-muestreo y permutaciones, para evaluar cambios tanto en la estructura global de red como en las conexiones de variables individuales o subgrupos de éstas. Este método esta implementado completamente en el paquete dna (differential network analysis) de R.

Otros métodos propuestos para la comparación de matrices de correlación o para matrices de precisión (la matriz de correlación invertida) incluyen los propuestos por

Ren et al. (Ren *et al.*, 2015), Belilovsky et al (Belilovsky, Varoquaux and Blaschko, 2016), Städler & Mukherjee (Städler and Mukherjee, 2017), Janková y van de Geer (Janková and van de Geer, 2017), y Xia y Li (Xia and Li, 2017), los cuales no garantizan ni estudian el control de la tasa de falsos positivos, por lo que no pueden controlar la tasa de error tipo I, así como tampoco estudian el comportamiento del error tipo II.

A pesar de su utilidad, los métodos mencionados anteriormente no proporcionan tratamiento especial para datos de naturaleza composicional. Además, algunos de estos métodos operan bajo suposiciones respecto a la distribución estadística de los datos y carecen de una caracterización de desempeño adecuada en términos de error de tipo I y poder estadístico para amplias regiones de dimensionalidad de los datos, como se puede apreciar de manera resumida en la Tabla 2.

3. DISEÑO DE UNA TÉCNICA PARA LA DISCRIMINACIÓN CUANTITATIVA DE ESTRUCTURAS DE CORRELACIÓN

En este capítulo se presentan, en primer lugar, dos técnicas no paramétricas existentes en la literatura que proveen pruebas de hipótesis para la evaluación de la similitud entre estructuras de correlación (sección 3.1). Estas pruebas son la de Krzanowski (Krzanowski, 1993) y la del paquete dna de R (Gill, Datta and Datta, 2014). La elección de estas dos técnicas como referencias para comparar la técnica propuesta obedece a las siguientes razones: i) No requieren el cumplimiento de supuestos de normalidad; ii) Los códigos/paquetes para su implementación se encontraron y se pudieron ejecutar sin inconvenientes; iii) Son métodos no paramétricos basados en permutaciones, lo que los hace directamente comparables con el método propuesto. Finalmente, se introduce la técnica propuesta en esta investigación en la sección 3.2.

3.1. Técnicas Existentes para la Discriminación de Matrices de Correlación

En esta sección se describen dos técnicas de la literatura para la comparación de estructuras de correlación. Éstas se usarán posteriormente para comparar su desempeño con el de la técnica propuesta en términos de error tipo I y potencia.

3.1.1. Prueba de Krzanowski

Krzanowski (Krzanowski, 1993) propuso un prueba no paramétrica basada en permutaciones para la evaluación de similitud entre matrices de correlación. Sean $Y_1 \in \mathbb{R}^{n_1 \times m}$ y $Y_2 \in \mathbb{R}^{n_2 \times m}$ las matrices de expresión (o conteos), y sean $D_1, D_2 \in \mathbb{R}^{m \times m}$ sus correspondientes matrices de correlación. El estadístico propuesto por Krzanowski es:

$$\mu = \mathrm{tr}(|D_1 - D_2|) \quad (1)$$

donde tr es la traza. Para la prueba de la hipótesis nula H_0: $D_1 = D_2$, las matrices de expresión son concatenadas en $Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \in \mathbb{R}^{(n_1+n_2)\times m}$. Luego, un gran número de permutaciones (e.g., $n_{perm} = 5\,000$) en las filas de Y son ejecutadas de manera que, en cada permutación, las matrices $Y_1^{\pi_i} \in \mathbb{R}^{n_1\times m}$ y $Y_2^{\pi_i} \in \mathbb{R}^{n_2\times m}$ son aleatoriamente seleccionadas y sus respectivas matrices de correlación $D_1^{\pi_i}$ y $D_2^{\pi_i}$ son calculadas. Luego, el estadístico $\mu^{\pi_i} = \mathrm{Tr}\left(abs\left(D_1^{\pi_i}, -D_2^{\pi_i}\right)\right)$ también es calculado. Finalmente, el valor *p* es calculado como sigue:

$$p_{\text{value}} = \frac{1}{n_{perm}} \sum_{i=1}^{n_{perm}} I(\mu^{\pi_i} > \mu) \qquad (2)$$

donde $I(\cdot)$ es una función indicadora.

3.1.2. Prueba dna

Gill et al. (Gill, Datta and Datta, 2010) propusieron el método dna junto con un paquete en R con el mismo nombre (Gill, Datta and Datta, 2014). Este paquete proporciona, entre otras opciones, una herramienta para la evaluación de las diferencias en la estructura de las redes de asociación de dos grupos con diferentes entornos biológicos (por ejemplo, control frente a pacientes enfermos). La técnica dna está diseñada para analizar redes (networks) construidas a partir de matrices $X_1 \in \mathbb{R}^{N_1\times p}$ y $X_2 \in \mathbb{R}^{N_2\times p}$ que contengan datos de niveles de expresión para p genes en N_1 y N_2 muestras, respetivamente. Sin embargo, esta técnica tiene un campo de aplicación amplio y es adecuada para datos de distinta naturaleza, tal como datos de abundancia de microbiota, por ejemplo.

Antes de explicar cómo funciona dna, es necesario proporcionar la definición de *módulo*. De acuerdo con Gill et al. (Gill, Datta and Datta, 2010), un módulo con un

parámetro de tamaño mínimo m y un parámetro de umbral de conectividad ϵ es un conjunto de genes, $\mathcal{F}$, que satisface las siguientes condiciones

- Su cardinalidad, $f = |\mathcal{F}|$, es al menos m.
- Dados dos genes, f_1 and f_2, que pertenecen a $\mathcal{F}$, están conectados por una trayectoria de $k \geq 2$ genes en $\mathcal{F}$, de manera que el valor mínimo de conectividad, $s_{1\,2}$, es mayor o igual que ϵ.

La conectividad puede ser medida con diferentes métricas; correlación de Pearson, correlación parcial, mínimos cuadrados parciales, entre otras. Para la evaluación de desempeño (ver Capítulo 4) de las técnicas presentadas en las secciones 3.1 y 3.2, se usará la correlación de Pearson como la métrica de conectividad.

Ahora, sea $\mathcal{M}_\kappa = \{\mathcal{F}_{k1}, \quad \cdots, \quad \mathcal{F}_{\kappa J_\kappa}\}$ una colección de todos los módulos J_κ contenidos en una red (network) κ, y sea $\mathcal{G}_0 = \bigcap_\kappa \bigcup_j \mathcal{F}_{\kappa j} \quad (j = 1, \dots, J_\kappa)$ la colección de todos los genes $(g_1, g_2, \dots, g_n)$ que pertenecen a (al menos) uno de los módulos. El estadístico propuesto por Gill et al. se calcula como sigue:

$$\mathcal{N} = 1 - \frac{1}{|\mathcal{G}_0|} \sum_{g \in \mathcal{G}_0} \frac{\left|\mathcal{F}_{1j}(g) \cap \mathcal{F}_{2j}(g)\right|}{\left|\mathcal{F}_{1j}(g) \cup \mathcal{F}_{2j}(g)\right|} \quad (3)$$

Un procedimiento de permutaciones aleatorias es llevado a cabo para evaluar significancia del estadístico calculado. Primero, las matrices X_1 y X_2 son concatendas en:

$$\mathbb{x} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \in \mathbb{R}^{(N_1+N_2)\times p} \quad (4)$$

Luego, un gran número de permutaciones (e.g., $n_{perm} = 5\,000$) de las filas de $\mathbb{x}$ llevado a cabo. En cada permutación, una nueva matriz, $\mathbb{x}^{\pi_i}$, de $(N_1 + N_2) \times p$ es obtenida. Sea $\mathbb{x}_1^{\pi_i}$ la matriz que contiene las primeras N_1 filas de $\mathbb{x}^{\pi_i}$, y sea $\mathbb{x}_2^{\pi_i}$ la matriz

que contiene sus últimas N_2 filas. El correspondiente estadístico, $\mathcal{N}^{\pi_i}$, es calculado usando la ecuación (3). El valor *p* es calculado como sigue:

$$p_{\text{value}} = \frac{1}{n_{boot}} \sum_{i=1}^{n_{boot}} I(\mathcal{N}^{\pi_i} \geq \mathcal{N}) \quad (5)$$

donde $I(\cdot)$ es una función indicadora.

3.2. Técnica Propuesta

En esta sección se presenta la técnica propuesta para la evaluación de similitud de matrices de correlación. La implementación de esta técnica fue llevada a cabo en el software R, y se encuentra adjunta en el Anexo A.1.

Primero, se explica el pretratamiento de datos propuesto para preparar los datos antes de llevar a cabo inferencias sobre las estructuras de correlación usando estadística convencional. Este pretratamiento integra un algoritmo de reemplazo de conteos nulos, la transformación CLR de Aitchison (Aitchison, 1982), y un auto-escalamiento de datos para eliminar sesgos por magnitudes dispares entre las variables. Posteriormente, se introduce un nuevo estadístico para la estimación de la desviación entre las estructuras de correlación de dos conjuntos de datos. Finalmente, se introduce un procedimiento basado en bootstrapping para la evaluación de significancia estadística en las desviaciones estimadas con el estadístico.

3.2.1. Pretratamiento de Datos

Sean $X_c^{\rho} \in \mathbb{R}^{n_c \times m}$ y $X_v^{\rho} \in \mathbb{R}^{n_v \times m}$ las matrices de datos de expresión para los controles y los casos (e.g., pacientes con alguna patología) respectivamente. Estas matrices contienen información sobre los niveles de expresión de m genes para n_c muestras de

control y n_v muestras caso. Ahora, se construye $X_T^\rho = \begin{bmatrix} X_c^\rho \\ X_v^\rho \end{bmatrix} \in \mathbb{R}^{(n_c+n_v)\times m}$, matriz que contiene todas las muestras.

Como se ha discutido previamente, cuando se tienen muestras de naturaleza composicional la transformación CLR (Aitchison, 1982) debe ser aplicada antes de estimar correlaciones. Sin embargo, antes de aplicar esta transformación, debe considerarse la presencia de conteos nulos, los cuales pueden resultar de muestras insuficientemente grandes o inexistentes. Debido a que la transformación CLR requiere que los conjuntos de datos contengan valores exclusivamente positivos, el uso de algún método para reemplazar conteos nulos es imperativo.

Se usa el algoritmo Bayesiano-Multiplicativo (BM) para reemplazo de ceros propuesto por Martín-Fernández et al. (Martín-Fernández *et al.*, 2015). Sea $\boldsymbol{x}_{T_i}^\rho \in \mathbb{R}^{1\times m}$ $(i = 1,2,\dots,n_c+n_v)$ la i-ésima fila de la matriz X_T^ρ. Entonces, el algoritmo BM remplaza los conteos nulos por:

$$BM\left(x_{T_{i,j}}^\rho\right) = \begin{cases} t_{i,j}\left(\frac{s_i}{n+s_i}\right), & \text{if } x_{T_{i,j}}^\rho = 0 \\ x_{T_{i,j}}^\rho\left(1 - \sum_{\forall k\,|x_{T_{i,j}}^\rho=0} t_{i,k}\left(\frac{s_i}{n+s_i}\right)\right), & \text{if } x_{T_{i,j}}^\rho \neq 0 \end{cases} \quad (6)$$

donde $n = \sum_{j=1}^m x_{T_{i,j}}^\rho$, $t_{i,j} = m^{-1}$ y $s_i = m$. Sea $X_T^{BM} \coloneqq BM\left(X_T^\rho\right)$ la matriz resultante luego de la aplicación del algoritmo BM, fila por fila, a la matriz X_T^ρ.

A continuación, con el fin de asegurar la composicionalidad de los datos, se aplica la operación *closure* (Aitchison, 1982) a cada fila de X_T^{BM}:

$$c\left(\boldsymbol{x}_{T_i}^{BM}\right) = \frac{k}{\sum_{j=1}^m x_{T_{i,j}}^{BM}}\, \boldsymbol{x}_{T_i}^{BM} \quad (7)$$

donde k es una constante arbitraria (usualmente $k = 100$). Sea $X_T^{BM,c}$ la matriz (concatenada) de datos luego de la aplicación del algoritmo BM y la operación (por

filas) de la operación *closure*. Para cualquier vector composicional $\boldsymbol{x} \in \mathbb{R}^D$ la transformación CLR está dada por:

$$clr(\boldsymbol{x}) = \left[\ln\frac{x_1}{g(\boldsymbol{x})},\ \ln\frac{x_2}{g(\boldsymbol{x})},\ \ldots,\ \ln\frac{x_D}{g(\boldsymbol{x})}\right] \quad (8)$$

donde $g(\boldsymbol{x}) = (\prod_{i=1}^{D} x_i)^{1/D}$ es la media geométrica. Aplicando la transformación CLR a $X_T^{BM,c}$ se obtiene finamente $X_T = \begin{bmatrix} X_c \\ X_v \end{bmatrix} \in \mathbb{R}^{(n_c+n_v)\times m}$. Luego, se divide la matriz X_T la matriz de controles ($X_c \in \mathbb{R}^{n_c \times m}$) y casos ($X_v \in \mathbb{R}^{n_v \times m}$). Entonces, la matriz de correlación de Pearson es calculada (Li and Ji, 2005):

$$S_c = \frac{1}{n_c - 1} X_{c_{norm}}^T X_{c_{norm}}, \qquad S_v = \frac{1}{n_v - 1} X_{v_{norm}}^T X_{v_{norm}} \quad (9)$$

donde $X_{g_{norm}} = \left(X_g - I_{n_g} b_g^T\right) \Sigma_g^{-1}$, $b_g = \frac{1}{n_g}\left(X_g\right)^T I_{n_g}$ contiene las medias de todas las m variables para el grupo g (controles o casos), $I_{n_g} = [1\ 1\ \ldots\ 1]^T \in \mathbb{R}^{n_g}$, y Σ_g es una matriz diagonal que contiene las desviaciones estándar de las m variables para el grupo g.

3.2.2. Estadístico para Estimar la Desviación de Matrices de Correlación

Sean S_c y S_v las matrices de correlación para X_c and X_v, respectivamente. La descomposición espectral para estas matrices es:

$$S_c = V_c \Lambda_c V_c^T, \qquad S_v = V_v \Lambda_v V_v^T \quad (10)$$

donde

$$\Lambda_c = \begin{bmatrix} \lambda_{c_1} & & \\ & \ddots & \\ & & \lambda_{c_m} \end{bmatrix} \in \mathbb{R}^{m\times m}, \quad \Lambda_v = \begin{bmatrix} \lambda_{v_1} & & \\ & \ddots & \\ & & \lambda_{v_m} \end{bmatrix} \in \mathbb{R}^{m\times m} \quad (11)$$

son matrices diagonales que contienen los autovalores para S_c and S_v, respectivamente. Sean $V_c = [\boldsymbol{v}_{c_1} \quad \boldsymbol{v}_{c_2} \quad \cdots \quad \boldsymbol{v}_{c_m}] \in \mathbb{R}^{m \times m}$ y $V_v = [\boldsymbol{v}_{v_1} \quad \boldsymbol{v}_{v_2} \quad \cdots \quad \boldsymbol{v}_{v_m}] \in \mathbb{R}^{m \times m}$ las matrices de autovectores de S_c y S_v.

Se introduce a continuación un nuevo estadístico para caracterizar las desviaciones en la estructura de correlación subyacente en los conjuntos de datos. Este estadístico primero requiere una reducción dimensional, i.e., la selección de los componentes principales para cada grupo de muestras. Este procedimiento, que está integrado al algoritmo e *análisis de componentes principales* (PCA) (Russell, Chiang and Braatz, 2012), consiste en determinar el número mínimo de autovalores a_c or a_v (para el grupo de controles o casos, respectivamente) que explica el $100\,(1-\alpha)$ % de la varianza total, i.e.:

$$\frac{\sum_{i=1}^{a_c} \lambda_{c_i}}{\sum_{i=1}^{m} \lambda_{c_i}} \leq (1-\alpha), \qquad \frac{\sum_{i=1}^{a_v} \lambda_{v_i}}{\sum_{i=1}^{m} \lambda_{v_i}} \leq (1-\alpha) \quad (12)$$

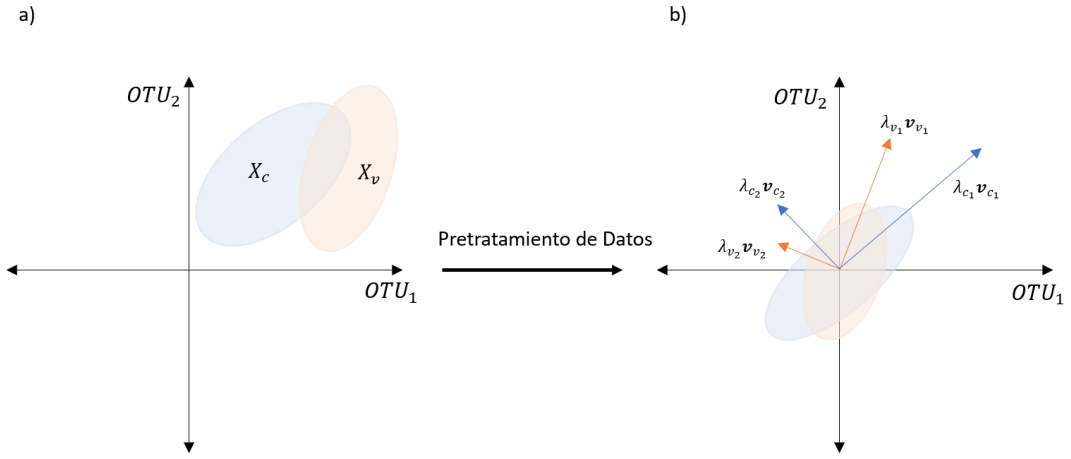

Figura 2 – Ilustración del efecto del pretratamiento de datos (izquierda), y de la proyección de los autovalores principales en las direcciones principales (derecha) para el cálculo del estadístico propuesto.

el efecto del pretratamiento de datos para X_c y X_v. La Figura 2 también muestra los autovectores (o direcciones principales) escalados por sus respectivos autovalores. El

estadístico propuesto busca medir las desviaciones angulares y de magnitud entre los vectores $\lambda_{c_j}\boldsymbol{v}_{c_j}$ y $\lambda_{v_j}\boldsymbol{v}_{v_j}$ (para $j = 1, \ldots, \max(a_c, a_v)$), y se define como sigue:

$$\varphi = \sum_{j=1}^{\max(a_c,a_v)} \left[\max\left\{\lambda_{c_j}, \lambda_{v_j}\right\}\left(\lambda_{c_j} - \lambda_{v_j}\right)\cos^{-1}\left(\boldsymbol{v}_{c_j}^T\boldsymbol{v}_{v_j}\right)\right] \quad (13)$$

donde $(\lambda_{c_j} - \lambda_{v_j})$ es la desviación de magnitud de los j-ésimos autovalores en Λ_c y Λ_v; $\cos^{-1}\left(\boldsymbol{v}_{c_j}^T\boldsymbol{v}_{v_j}\right)$ calcula la desviación angular entre los j-ésimos autovectores en V_c y V_v; y $\max\left\{\lambda_{c_j}, \lambda_{v_j}\right\}$ proporciona un factor de ponderación, de manera que la contribución de la j-ésima desviación al índice φ es proporcional a la importancia relativa de los componentes principales.

3.2.3. Prueba de Significancia basada en Bootstrapping para Inferencia de Igualdad de Estructuras de Correlación

La Figura 3 muestra un diagrama de flujo ilustrando todos los pasos necesarios para la implementación de la técnica basada en bootstrapping (Efron and Tibshirani, 1986; Kohavi and others, 1995) para hacer inferencia sobre φ y calcular significancia estadística de la desviación calculada.

El procedimiento comienza con $X_T = \begin{bmatrix} X_c \\ X_v \end{bmatrix} \in \mathbb{R}^{(n_c+n_v)\times m}$, calculándose el estadístico de prueba, φ, usando la ecuación (13). Denótese como φ_{act} el valor calculado para φ con los datos reales. El siguiente paso consiste en la selección aleatoria (con repetición de filas permitida) de dos submuestras, $X_c^{*i} \in \mathbb{R}^{n_c\times m}$ y $X_v^{*i} \in \mathbb{R}^{n_v\times m}$, de X_T en cada iteración de Bootstrap i $(i = 1,2, \ldots, n_{boot})$, donde n_{boot} es el número de submuestreos (e.g., $n_{boot} = 5\,000$). Las submuestras se guardan en:

$$\mathcal{Z}_{c,v}^{*} = \begin{Bmatrix} X_c^{*1} & X_v^{*1} \\ X_c^{*2} & X_v^{*2} \\ \vdots & \vdots \\ X_c^{*n_{boot}} & X_v^{*n_{boot}} \end{Bmatrix} \quad (14)$$

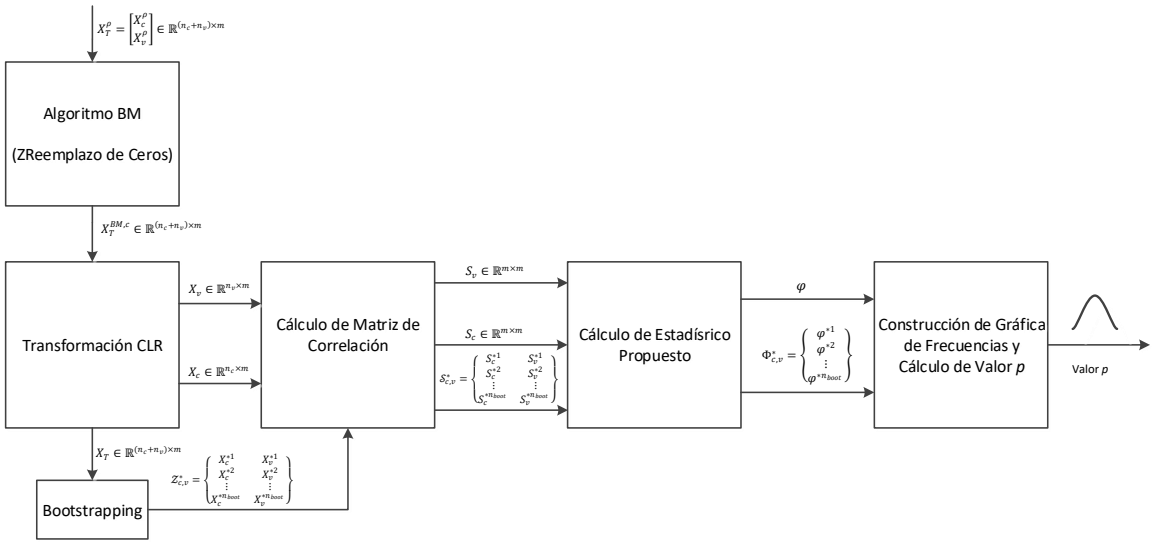

Figura 3 – Diagrama de flujo de la técnica propuesta para la evaluación de similitud de estructuras de correlación.

Nótese que para cada submuestra las matrices seleccionadas tienen las mismas dimensiones de X_c and X_v. A continuación, se calcula el estadístico φ para cada submuestra, resultando en el vector $\boldsymbol{\Phi}_{\boldsymbol{c,v}}^{*} = (\varphi^{*1}, \varphi^{*2}, \ldots, \varphi^{*n_{boot}})$. A continuación, se construye un histograma de frecuencia para $\Phi_{c,v}^{*}$ y se localiza φ_{act} en el eje x, representándolo como una línea vertical roja, como se ilustra en la Figura 4.

Finalmente, el valor *p* se calcula como sigue:

$$p_{\text{value}} = \frac{1}{n_{boot}} \sum_{i=1}^{n_{boot}} I(|\varphi^{*i}| > |\varphi_{act}|) \quad (15)$$

donde $I(\cdot)$ es una función indicadora.

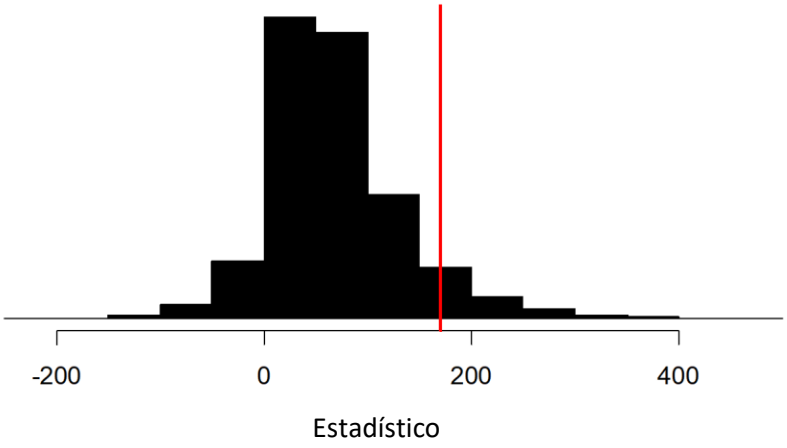

Figura 4 – Histograma de frecuencia para el estadístico propuesto. φ_{act} se muestra como línea vertical roja.

3.3. EVALUACIÓN DE DEMPEÑO DE LAS TÉCNICAS DE COMPARACIÓN DE ESTRUCTURAS DE CORRELACIÓN

En sección se describe el experimento realizado, con datos sintéticos, para la evaluación del desempeño de la técnica propuesta y de las dos técnicas descritas en la sección 3.1 para la comparación de estructuras de correlación entre dos conjuntos de datos.

3.3.1. Evaluación de Error Tipo I y Potencia Estadística

Se compara el desempeño de las pruebas de Krzanowski, dna y la propuesta, discutidas en sección 3.1, mediante el cálculo empírico de la tasa de error tipo I (TIER, por sus siglas en inglés) y la potencia estadística cuando se comparan dos matrices de correlación. Para este propósito, se desarrolló un experimento, cuya implementación en R se encuentra disponible en el Anexo A.2, y que comprende los siguientes tres pasos resumidos esquemáticamente en la Figura 5:

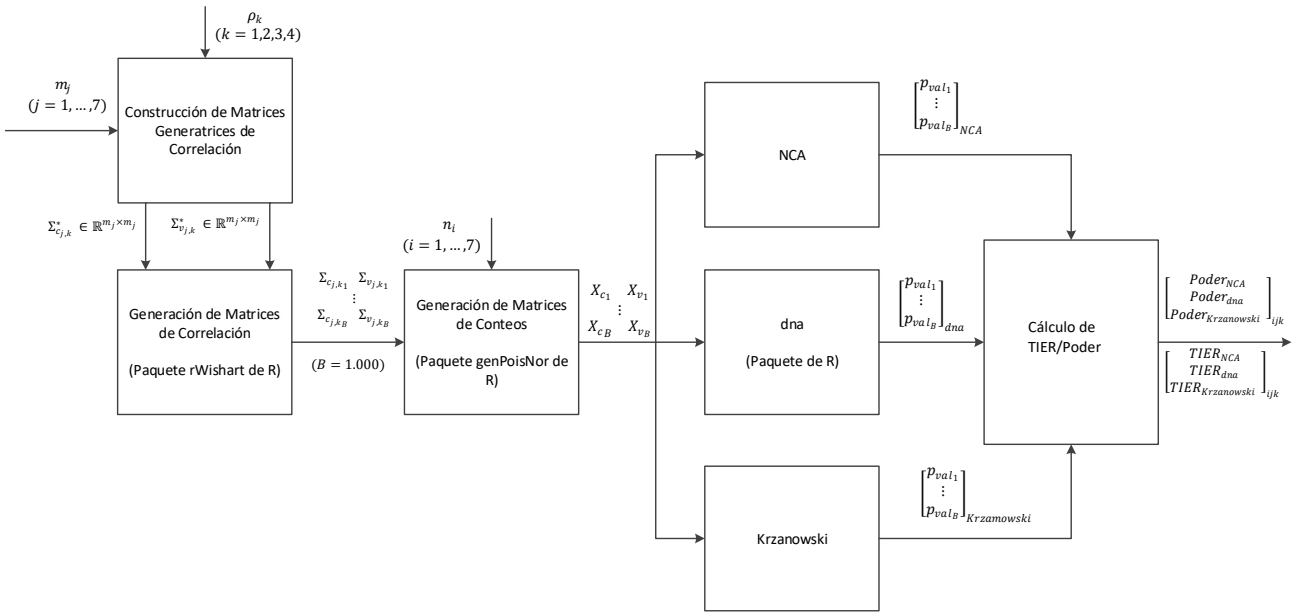

Figura 5 – Diagrama de flujo del experimento para a evaluación de error tipo I y potencia estadística.

Paso 1: Generación de Datos

Esta etapa comprende a su vez los siguientes pasos:

1. Se define la tripleta (n_i, m_j, ρ_k). Se establece: $n = \{20, 60, 100, 140, 180, 220, 260\}$, $m = \{20, 40, 60, 80, 100, 120, 140\}$, $\rho = \{0, 0.1, 0.5, 0.9\}$.

2. Para cada tripleta (n_i, m_j, ρ_k), se construye un par de matrices de correlación generatrices, $\Sigma_{c_{j,k}}$ y $\Sigma_{v_{j,k}}$, de manera que $\Sigma_{c_{j,k}} = I_{m_j}$ y $\Sigma_{v_{j,k}} = (1-\rho)I_{m_j} + \rho 1_{m_j} 1_{m_j}^T$, donde $I_{m_j} \in \mathbb{R}^{m_j \times m_j}$ es la matriz identidad y $1_{m_j} \in \mathbb{R}^{m_j \times 1}$ es un vector columna de unos. Este procedimiento es similar al llevado a cabo por Vélez & Correa (Vélez and Correa, 2013).

3. Para cada par de matrices de correlación generatrices $\Sigma_{c_{j,k}}$ y $\Sigma_{v_{j,k}}$, respectivamente, se generan *B* pares de matrices X_{c_r} y X_{v_r} (*r*=1,2,…,*B*) de dimensión $n_i \times m_j$, cuyos datos siguen una distribución multivariada de Poisson, utilizando el paquete genPoisNor de R (Amatya and Demirtas, 2017). La elección de la distribución de Poisson como modelo estadístico subyacente para la generación de datos se basa en el hecho de que, en el análisis de expresión de genes y conteos de OTUs, los datos representan recuentos (positivos) de secuencias (Nguyen *et al.*, 2016; Xia and Sun, 2017). El número de réplicas experimentales fue *B*=1 000.

Paso 2: Aplicación de las pruebas estadísticas

Para cada (n_i, m_j, ρ_k, r), se fija el número de réplicas de bootstrap (o permutaciones, dependiendo del método usado) como $n_{boot} = n_{perm} = 5\,000$ para las pruebas de Krzanowski, dna y la propuesta. Para la implementación de la prueba de dna se usó la correlación de Pearson como la métrica de interacción y se usó la configuración predeterminada para los demás parámetros. Finalmente, se guardan los valores *p* obtenidos para cada prueba.

Paso 3: Estimación de la TIER y la Potencia Estadística

Para aquellas condiciones experimentales en las que $\rho = 0$, la matriz de correlación es igual para cada par de matrices X_{c_r} y X_{v_r} (*r*=1,2,..,*B*), y la TIER empírica es calculada como la proporción de veces, en las *B* réplicas, que la hipótesis nula H_0 es rechazada (esto es, que se determina que las matrices de correlación de los dos grupos comparados son estadísticamente diferentes) para una probabilidad nominal de error tipo I preespecificada α. Nótese, por otra parte, que para cualquier $\rho \neq 0$ las matrices de correlación $\Sigma_{c_{j,k}}$ y $\Sigma_{v_{j,k}}$ son diferentes por definición, y la tasa de rechazo de H_0,

calculada para cada combinación $(n_i, m_j, p_k \neq 0)$, corresponde a la potencia estadística de cada prueba. Se usa $\alpha = 0.05$ para todos los casos.

3.3.2. Resultados de Evaluación de Desempeño

La Figura 6a muestra las superficies de respuesta de la TIER obtenidas para las pruebas evaluadas. En general, la técnica propuesta exhibe un mejor desempeño que las pruebas de Krzanowski y dna, sugiriendo que la prueba propuesta controla eficientemente la probabilidad de error tipo I al nivel preespecificado sin importar el tamaño de la muestra ni el número de variables. Es notable que la prueba de Krzanowski exhibe TIERs por encima del nivel nominal de 5%. Las Figuras 6b, 6c y 6d muestran los resultados para ρ=0.1, 0.5 and 0.9.

Independientemente del valor de ρ ($\rho \neq 0$) y de la dimensión de los datos, la técnica propuesta y la de Krzanowski exhibieron tasas de rechazo altas y similares. Por otra parte, la prueba de dna se desempeña pobremente comparada con las dos primeras. En términos generales, la técnica propuesta controla eficientemente la probabilidad de error tipo I y exhibe una buena potencia estadística en los escenarios evaluados, superando a las otras dos técnicas evaluadas.

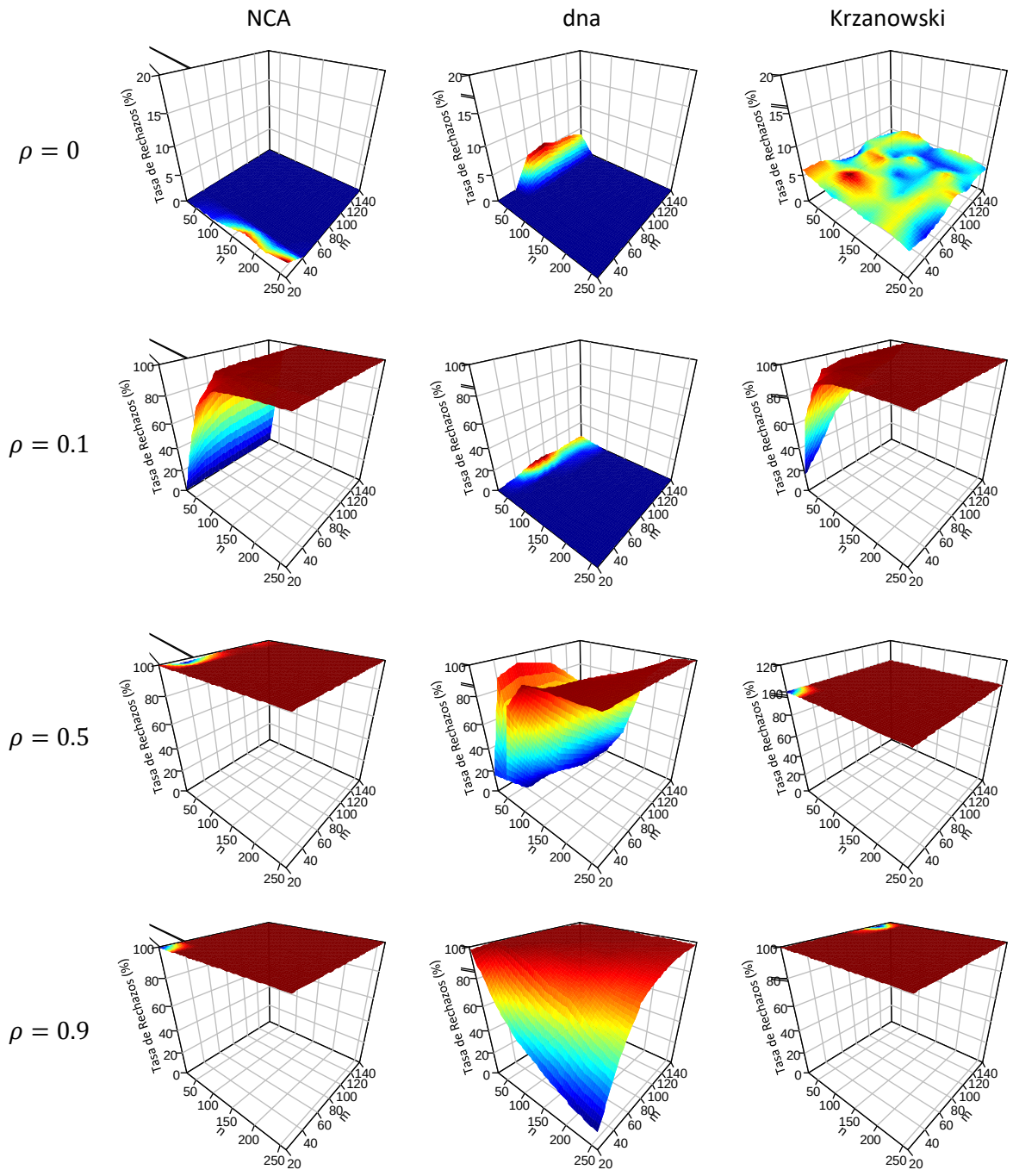

Figura 6 – Tasa de rechazos de la hipótesis nula (%) en función del tamaño de muestra (n) y el número de variables (m) cuando ρ=0, ρ=0.1, ρ=0.5 y ρ=0.9.

4. DISEÑO DE LA TÉCNICA DE ANÁLISIS DIFERENCIAL

En esta sección se desarrolla una técnica de análisis diferencial que permite identificar aquellas variables (e.g., OTUs) que más han contribuido a cambiar la estructura de correlación al comparar dos conjuntos de datos.

Supóngase que el procedimiento de comparación de matrices de correlación, diseñado en el capítulo 3 (sección 3.2), revelase que los conjuntos de datos de casos y controles, X_c y X_v, exhiben diferencias significativas en sus estructuras de correlación. En lugar de buscar variables (e.g., genes, OTUs) cuya abundancia relativa haya cambiado diferencialmente, en este capítulo se propone un método para identificar aquellas variables que son más probablemente responsables por la diferenciación en las estructuras de correlación, esto es, aquellas variables que han cambiado significativamente sus interacciones en la red al cambiar de una condición biológica a la otra. Los códigos en Matlab para la implementación de esta técnica se adjuntan en el anexo A.3. Inicialmente, tómese el conjunto de muestras de control, X_c, para ejecutar la extracción de características (features) de la técnica de PCA convencional (Russell, Chiang and Braatz, 2012), como se describe más adelante.

La estructura propia (o *eigenstructure*, en inglés) de este conjunto de entrenamiento es determinada usando las ecuaciones (10) y (11), obteniendo Λ_c y V_c. Además, una reducción dimensional es llevada a cabo usando la ecuación (12), obteniendo el número de componentes principales, a_c. A continuación, se construye la matriz de carga (o *loading matrix*, en inglés) $P \in \mathbb{R}^{m \times a_c}$ reteniendo únicamente las primeras a_c columnas de Λ_c, y se calcular la matriz de puntuaciones (o *score matrix*, en inglés) $T \in \mathbb{R}^{n_c \times a_c}$ como $T = X_c P_c$. Sea $\Gamma \in \mathbb{R}^{m \times m}$ una matriz tal que $\Lambda_c = \Gamma^T \Gamma$, y sea $\Gamma_a \in \mathbb{R}^{a_c \times a_c}$ una matriz que contiene las primeras filas y columnas de Γ. Con el fin de llevar a cabo el análisis estadístico, el grupo de muestras de controles ha sido usado de manera análoga al *conjunto de entrenamiento* (o *training set*, en inglés) en las aplicaciones tradicionales de PCA para la detección y diagnóstico de fallas en procesos industriales,

tal como es explicado por (Jeng, 2010; Portnoy *et al.*, 2016). De manera similar, el grupo de muestras de casos, X_v, es equiparado al *conjunto de prueba* (o *testing set*, en inglés) en PCA. En consecuencia, para cada vector (fila) en el conjunto de prueba, $\boldsymbol{x} \in \mathbb{R}^{1\times m}$, el estadístico T^2 de Hotelling es calculada para esta muestra así:

$$T^2 = \boldsymbol{x} P \Gamma_a^{-2} P^T \boldsymbol{x}^T \quad (16)$$

El umbral de detección para el estadístico T^2 se calcula como sigue:

$$T_\alpha^2 = \frac{a_c(n_c-1)(n_c+1)}{n_c(n_c-a_c)} F_\alpha(a_c, n_c - a_c) \quad (17)$$

donde $F_\alpha(a_c, n_c - a_c)$ es la función de distribución acumulativa inversa de Fisher para un nivel de significancia de α. Luego, el estadístico Q se calcula como:

$$Q = \boldsymbol{r}^T \boldsymbol{r} \quad (18)$$

donde $\boldsymbol{r} = (I - PP^T)\boldsymbol{x}^T$, y el umbral de detección para Q está dado por:

$$Q_\alpha = \theta_1 \left[\frac{h_0 c_\alpha \sqrt{2\theta_2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right]^{1/h_0} \quad (19)$$

con $\theta_i = \sum_{j=a+1}^{n} \sigma_j^{2i}$, $h_{0_{k+1}} = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}$, $\sigma_i = \sqrt{\lambda_i}$, y c_α es la desviación normal correspondiente al nivel de confianza $(1-\alpha)100\%$. La Figura 7 muestra un ejemplo de dos cartas de control con los estadísticos T^2 de Hotelling y Q para cuatro diferentes muestras (i.e., pacientes).

Miller et al. (Miller, Swanson and Heckler, 1998) propusieron un procedimiento, muestra-a-muestra, para estimar la contribución de cada variable x_j $(j = 1, \dots, m)$ al estado de 'fuera de control' cuando sea que ocurra una violación de alguno de los umbrales de detección. El cálculo de contribución contiene los siguientes pasos:

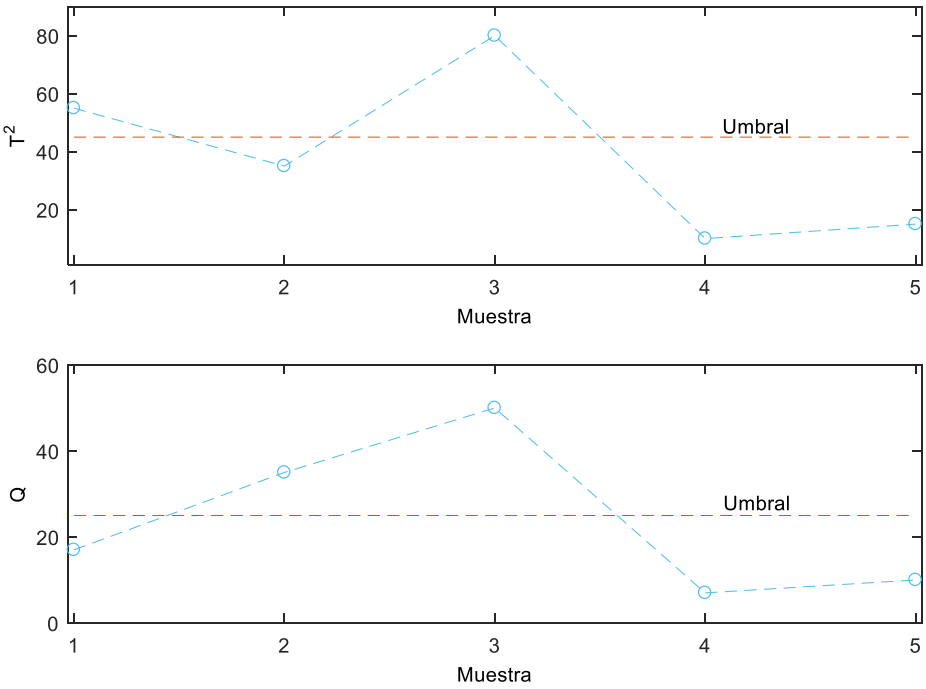

Figura 7 – Cartas de control para los estadísticos T^2 de Hotelling y Q.

1. Verificar las *puntuaciones* normalizadas $\frac{t_i}{\sigma_i}$ $(i = 1, \dots, m)$, donde $\boldsymbol{t}$ es la j-ésima columna de la matriz de puntuaciones T. Almacenar las puntuaciones que satisfagan $\frac{t_i}{\sigma_i} > (T_\alpha^2)^{1/a}$.
2. Determinar la contribución de cada variable x_j a las puntuaciones previamente guardadas como

$$cont_{i,j} = \frac{t_i}{\sigma_i^2} P_{i,j}(x_j - \mu_j) \quad (20)$$

donde los elementos $P_{i,j}$ son las entradas de la matriz de carga P. Si $cont_{i,j}$ es negativo, se hace $cont_{i,j} = 0$.

3. Finalmente, la contribución total de la variable x_j, para una muestra en particular, se calcula como

$$CONT_j = \sum_{i=1}^{r} cont_{i,j} \quad (21)$$

donde r es el número de puntuaciones retenidas en el paso 1.

Se propone ahora, para cada muestra $\boldsymbol{x}_{\boldsymbol{v}_k}$ $(k = 1, \dots, n_v)$, calcular la contribución de todas las variables x_j usando el procedimiento previamente explicado y posteriormente guardar estas contribuciones (muestra-a-muestra) en el vector $\mathfrak{c}_k$ dado por

$$\mathfrak{c}_k = [CONT_1^k \quad CONT_2^k \quad \cdots \quad CONT_m^k]^T \quad (22)$$

A continuación, guárdense todos estos vectores de contribución muestra-a-muestra en una *matriz de contribución*, $\mathfrak{C}$, como sigue:

$$\mathfrak{C} = [\mathfrak{c}_1 \quad \mathfrak{c}_2 \quad \cdots \quad \mathfrak{c}_{n_v}] \in \mathbb{R}^{m \times n_v} \quad (23)$$

Para $k = 1, \dots, n_v$, se definen los *factores de ponderación* (o *weighting factors*, en inglés) w_k, como

$$w_k = \begin{cases} \dfrac{T_k^2}{T_\alpha^2}, & \text{if } T_k^2 > T_\alpha^2 \text{ and } Q_k \leq Q_\alpha \\ \dfrac{Q_k}{Q_\alpha}, & \text{if } T_k^2 \leq T_\alpha^2 \text{ and } Q_k > Q_\alpha \\ \dfrac{T_k^2}{T_\alpha^2}\dfrac{Q_k}{Q_\alpha}, & \text{if } T_k^2 > T_\alpha^2 \text{ and } Q_k > Q_\alpha \\ 0, & \text{if } T_k^2 \leq T_\alpha^2 \text{ and } Q_k \leq Q_\alpha \end{cases} \quad (24)$$

los cuales dependen de que tan grandes son las violaciones de los umbrales, si las hay. El primer caso considerado en la ecuación (24) es ilustrado por la muestra 1 en la Figura 7, ya que únicamente el umbral T_α^2 es violado. La muestra 2 ilustra el segundo caso de la ecuación (24), con una violación exclusiva del umbral Q_α. La muestra 3 lustra el tercer caso, en el que ocurren violaciones de ambos umbrales. Finalmente, la muestra

4 ilustra el último caso, en el que no ocurra ninguna violación de umbrales y el factor de ponderación es 0.

Sea, ahora, $\boldsymbol{w} = [w_1 \quad w_2 \quad \cdots \quad w_k]^T$ un vector columna que contiene todos los factores de ponderación. Nótese que la matriz $\mathfrak{C}$ contiene todos los vectores de contribución muestra-a-muestra, y que cada uno de éstos contiene, para una muestra en particular, la importancia relativa de todas las variables respecto a su contribución a la diferenciación. Para cuantificar esta importancia relativa teniendo en cuenta todas las muestras del conjunto de prueba, se propone un *vector de contribución total ponderada* (o *overall weighted contribution vector*, en inglés), $\mathfrak{c}_{\boldsymbol{w}}$, que se calcula como

$$\mathfrak{c}_{\boldsymbol{w}} = c(\mathfrak{C} \cdot \boldsymbol{w}) \in \mathbb{R}^{1\times m} \quad (25)$$

donde $c(\cdot)$ es la operación closure previamente definida en la ecuación (7). Usando $k = 100$ como la constante de closure, las contribuciones totales ponderadas son proporcionadas como porcentajes, las cuales pueden ser mostradas en un histograma, tal como se ilustra en la Figura 8.

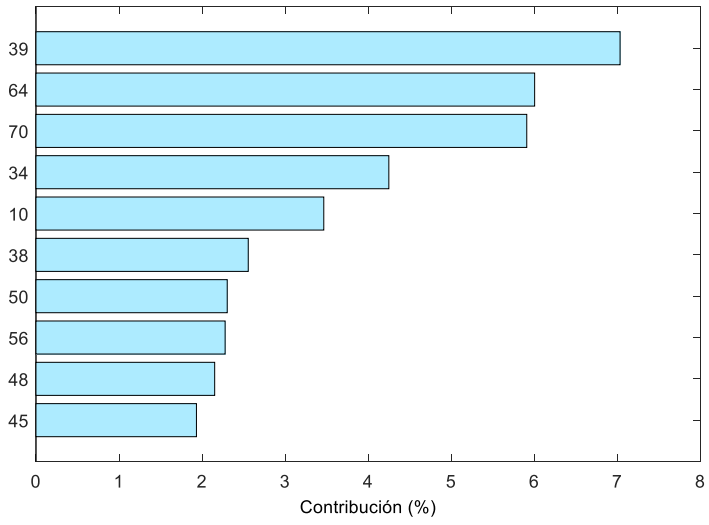

Figura 8 – Contribución de las variables individuales a la diferencia en las estructuras de correlación de los grupos comparados.

4.1. EVALUACIÓN DEL DESEMPEÑO DE LA TÉCNICA PROPUESTA PARA ANÁLISIS DIFERENCIAL

En este capítulo se describe el experimento realizado, con datos sintéticos, para la evaluación del desempeño de la técnica propuesta para el análisis diferencial. En la 4.1 se define un nuevo indicador de desempeño que se usara para la evaluación de los resultados del experimento propuesto, la sección 6.2 muestra el diseño del experimento y, finalmente, la sección 6.3 muestra los resultados obtenidos.

4.1.1. Indicador de Desempeño Propuesto Para la Técnica de Análisis Diferencial

Supóngase de nuevo que se tienen dos conjuntos de datos de controles y casos, X_c y X_v respectivamente, de los cuales se sabe con certeza que hay diferencias en sus estructuras de correlación. Además, supóngase que se sabe a priori cuáles son las p ($p < m$) variables que contribuyen significativamente a esas diferencias existentes. Se almacenan ahora las variables en orden de importancia según su contribución (de mayor a menor) en $\mathcal{C} = \begin{bmatrix} x_{\mathcal{C}_1} & \cdots & x_{\mathcal{C}_p} & \cdots & x_{\mathcal{C}_m} \end{bmatrix}$, de manera que:

$$cont\left(x_{\mathcal{C}_i}\right) \geq cont\left(x_{\mathcal{C}_j}\right), \qquad \forall\, i,j \in (1,2,\dots,m) \mid i < j \quad (26)$$

Si se aplicase la técnica propuesta para análisis diferencial y se obtuviese el vector de contribuciones ponderadas, $\mathbf{c_w} = \begin{bmatrix} cont\left(x_{\mathcal{C}_1}\right) & cont\left(x_{\mathcal{C}_2}\right) & \cdots & cont\left(x_{\mathcal{C}_m}\right) \end{bmatrix}$, idealmente se debería satisfacer la ecuación (26). Se propone, en la ecuación (27), un indicador para medir la efectividad de la técnica propuesta para la identificación de las variables más contribuyentes a la diferenciación de las estructuras de correlación:

$$AIE = \frac{100}{p}\sum_{i=1}^{p}\left[\frac{1}{m-i}\sum_{j=i+1}^{m} I\left(cont\left(x_{\mathcal{C}_i}\right) \geq cont\left(x_{\mathcal{C}_j}\right)\right)\right] \quad (27)$$

donde $I(\cdot)$ es una función indicadora. A este indicador se le denomina '*efectividad promedio de identificación (de variable*s)' o, en inglés, '*Average (variable)*

Identification Effectiveness' (AIE). Si la técnica propuesta identifica perfectamente las contribuciones de cada variable, de manera que se satisfaga siempre la ecuación (26), el valor del AIE será 100. Por otro lado, si la técnica errase completamente en la estimación de las contribuciones, asignándole mayores valores de contribución a aquellas variables con contribuciones despreciables o nulas y asignándole los menores valores a las variables más contribuyentes, el valor del AIE sería de cero (0). El AIE se utilizará como indicador para la evaluación de desempeño de la técnica propuesta para análisis diferencial, la cual se desarrolla en la siguiente sección.

4.1.2. Diseño del Experimento para Evaluación del Desempeño de la Técnica de Análisis Diferencial

Se describe aquí el experimento llevado a cabo para la evaluación del desempeño de la técnica propuesta para análisis diferencial. Este experimento comprende los pasos descritos a continuación:

Paso 1: Generación de Datos

Este paso comprende a su vez los siguientes pasos:

1. Se define la tripleta $\left(n_i, m_j, \sigma_k^2\right)$. Se establecen los siguientes niveles para los factores: $n = \{20, 60, 100, 140, 180, 220, 260\}$, $m = \{20, 40, 60, 80, 100, 120, 140\}$, $\sigma^2 = \{20, 50, 100\}$.

2. Para cada tripleta $\left(n_i, m_j, \sigma_k^2\right)$, se construye un par de matrices de covarianza generatrices, $\Psi_{c_{j,k}}$ y $\Psi_{v_{j,k}}$, de manera que $\Psi_{c_{j,k}} = I_{m_j} \in \mathbb{R}^{m_j \times m_j}$ es la matriz identidad. Por otro lado, $\Psi_{v_{j,k}}$ se construye a partir de I_{m_j}, pero haciendo $\sigma_{lq}^2 = \sigma_k^2$ para todo $l = (1,2,3)$ y $q = (1,2,3)$ tal que $l \neq q$, y para $l = q = 1$, donde σ_{lq}^2 representa las entradas de la matriz $\Psi_{v_{j,k}}$. De esta manera, se logra que el grupo de controles (c) exhiba una varianza de 1 para todas sus variables a la vez

que no se exhiben covarianzas (y por lo tanto tampoco correlaciones) entre las mismas. Por otro lado, se inducen cambios en la varianza de la primera variable y su covarianza con la segunda y la tercera variable del grupo de casos (v). Por la forma en la que se han definido las matrices de covarianza generatrices, se sabe a priori que la primera variable será la que más contribuya a las diferencias exhibidas en las estructuras de correlación entre el grupo de controles y casos. También se construye un vector de medias $b = [\bar{x}_1 \quad \bar{x}_2 \quad \cdots \quad \bar{x}_{m_j}]$. Para este experimento se dejaron todas las medias fijas e iguales a 100 para todos los escenarios evaluados.

3. Para cada par de matrices de covarianza generatrices, $\Psi_{c_{j,k}}$ y $\Psi_{v_{j,k}}$ respectivamente, se generan B pares de matrices X_{c_r} y X_{v_r} (r=1,2,…,B) de dimensión $n_i \times m_j$, cuyos datos siguen una distribución multivariada de Poisson, utilizando la función sampleCovPoisson de Matlab (Bethge and Berens, 2008; Macke *et al.*, 2009), la cual no está integrada al software pero se encuentra disponible en línea. sampleCovPoisson requiere como entradas el vector de medias de las variables, el número de muestras que se desea generar y la matriz de covarianza generatriz. El número de réplicas experimentales fue $B = 100$.

Paso 2: Aplicación de la técnica propuesta

Para cada $(n_i, m_j, \sigma_k^2, r)$, se aplica la técnica propuesta para análisis diferencial y se guarda el vector de contribución total ponderada, $\mathfrak{c}_{\boldsymbol{w}}$.

Paso 3: Estimación del indicador de desempeño AIE

Para cada una de las condiciones experimentales (n_i, m_j, σ_k^2), se calcula el AIE obtenido en las B=100 réplicas utilizando la ecuación (27) con $p = 1$, se promedian estos B valores y se reporta el resultado de este promedio como el valor de AIE para

dicha condición. Se guardan todos los resultados obtenidos en todas las condiciones experimentales.

4.1.3. Resultados de la Evaluación de Desempeño de la Técnica Propuesta para Análisis Diferencial

La Figura 9 muestra los resultados de la evaluación de desempeño de la técnica propuesta para análisis diferencial en términos de su efectividad promedio de identificación de variables (AIE) contribuyentes a la diferenciación.

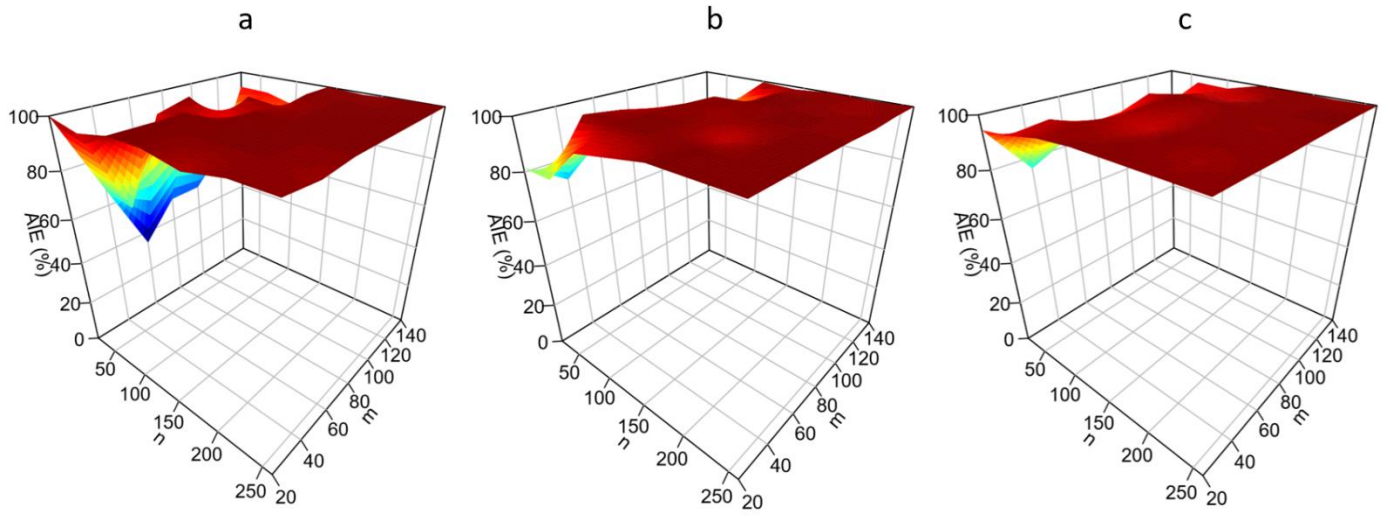

Figura 9 – Resultados experimentales de AIE (con $p = 1$) en función del número de muestras (n) y el número de variables (m) para: a) $\sigma^2 = 20$; b) $\sigma^2 = 50$; c) $\sigma^2 = 100$.

En términos generales, la técnica se desempeñó apropiadamente, ya que se observan, en casi todas las condiciones experimentales, valores altos de AIE. Además, es notable que la técnica se desempeñó mejor para valores grandes de m (número de variables), lo cual la hace apropiada para el tipo de datos provenientes de secuenciación, que suelen ser de gran dimensión.

5. DISEÑO DE UNA TÉCNICA DE REDUCCIÓN DIMENSIONAL PARA LA INTEGRACIÓN Y PONDERACIÓN DE NUEVAS MUESTRAS/PACIENTES

En este capítulo se propone una nueva técnica de reducción dimensional que permite la integración y ponderación de nuevas muestras/pacientes en la base de datos de uno de los grupos (controles o casos). En la sección 5.1 se aborda el escenario de la inclusión de una (sola) nueva muestra. En la sección 5.2 se aborda, por otro lado, el escenario de la inclusión de un nuevo conjunto de muestras.

5.1. Escenario I: Integración de una Nueva Muestra a una Base de Datos para la Determinación de la Distorsión Causada a su Estructura de Correlación

De nuevo, considérese que la información de los grupos de pacientes de control y casos se tiene almacenada en $X_c^{\rho} \in \mathbb{R}^{n_c \times m}$ y $X_v^{\rho} \in \mathbb{R}^{n_v \times m}$ respectivamente. Para el grupo g (ya sea controles, $g = c$, o casos, $g = v$), llévese a cabo el pretratamiento de datos descrito en la sección 3.2.1, el cual consiste en la aplicación (fila por fila) del algoritmo BM para reemplazo de ceros, la operación closure y la transformación CLR de Aitchison (Aitchison, 1982), como sigue:

$$X_g = clr\left(c\left(BM\left(X_g^{\rho}\right)\right)\right) \quad (28)$$

Adicionalmente, se aplica una normalización, como la hecha en la ecuación (9), obteniéndose:

$$X_{g_{norm}} = \left(X_g - I_{n_g} b_g^T\right)\Sigma_g^{-1} \quad (29)$$

donde $I_{n_g} = [1 \quad 1 \quad \cdots \quad 1]^T \mathbb{R}^{n_g \times 1}$, $b_g \mathbb{R}^{m \times 1}$ es un vector columna que contiene las medias de todas las variables en X_g, mientras que $\Sigma_g \mathbb{R}^{m \times m}$ es una matriz diagonal que

contiene las desviaciones estándar (σ_{g_i}, para $i = 1, \dots, m$) de todas las variables. Además, calcúlese la matriz de correlación de $X_{g_{norm}}$ como $S_g = \frac{1}{n_g - 1} X_{g_{norm}}^T X_{g_{norm}}$. Ahora supóngase que una nueva muestra, $\boldsymbol{x}_p^\rho \in \mathbb{R}^{1 \times m}$, correspondiente (por ejemplo) a un nuevo paciente p, es tomada. También debe hacerse el pretratamiento de datos para esta muestra, como sigue:

$$\boldsymbol{x}_p = clr\left(c\left(BM\left(\boldsymbol{x}_p^\rho\right)\right)\right) \quad (30)$$

Sería interesante evaluar la disrupción causada en la estructura de correlación del grupo g (controles o casos) por la incorporación de esta nueva muestra en su base de datos. La manera intuitiva de abordar la evaluación de esta disrupción consistiría en la integración de la nueva muestra, $\boldsymbol{x}_p$, en el conjunto X_g y (re)calcular la matriz de correlación para una posterior evaluación de su disrupción. Sin embargo, surge un problema al considerar que el conjunto de datos X_g podría contener muestras de un gran número de pacientes. En ese caso, una sola muestra podría no ser lo suficientemente *pesada* como para causar una disrupción significativa de la estructura de correlación, incluso si exhibiese una abundancia relativa de las variables diferente a aquella en las muestras que componen a X_g.

Un enfoque para abordar el problema de dimensionalidad anteriormente expuesto consiste en sub-muestrear al azar un número pequeño de filas de X_g y luego combinarlas con $\boldsymbol{x}_p$ para la evaluación posterior de su disrupción. Este enfoque, sin embargo, dejaría por fuera mucha información, contenida en las filas sin seleccionar durante el procedimiento de submuestreo. Se propone, en contraste, un enfoque novedoso para la reducción de la dimensionalidad que permite una evaluación equitativa (ponderada) de la disrupción de la estructura de correlación de un conjunto de datos X_g causada por la incorporación de una nueva muestra, $\boldsymbol{x}_p$, utilizando toda la información contenida en el conjunto de datos original.

Para lograr este enfoque, primero se encuentra una expresión para la matriz de correlación distorsionada que revela los pesos naturales de las contribuciones del conjunto de datos original, X_g, y la nueva muestra, $\boldsymbol{x}_p$, en la constitución de la nueva estructura de correlación. Una vez que se identifican estos pesos naturales, se introduce una modificación para proporcionar pesos más balanceados (ponderados) para la evaluación de la disrupción. Considérese entonces que se concatenan los datos en:

$$\tilde{X}_g = \begin{bmatrix} X_g \\ \boldsymbol{x}_p \end{bmatrix} \mathbb{R}^{\tilde{n}_g \times m} \quad (31)$$

donde $\tilde{n}_g = n_g + 1$ es el número de filas (muestras) del conjunto de datos concatenados $\tilde{X}_g$. La ecuación (31) puede ser combinada con la ecuación (29), obteniéndose:

$$\tilde{X}_g = \begin{bmatrix} X_{g_{norm}} \Sigma_g + I_{n_g} b_g^T \\ \boldsymbol{x}_p \end{bmatrix} \quad (32)$$

Antes de calcular la matriz de correlación distorsionada, $\tilde{S}_g$, se normaliza $\tilde{X}_g$ como sigue:

$$\tilde{X}_{g_{norm}} = \left(\tilde{X}_g - I_{\tilde{n}_g} \tilde{b}_g^T\right) \tilde{\Sigma}_g^{-1} = \begin{bmatrix} \left(X_{g_{norm}} \Sigma_g - I_{n_g} \Delta b_g^T\right) \tilde{\Sigma}_g^{-1} \\ \boldsymbol{x}_{p_{norm}} \end{bmatrix} \quad (33)$$

donde $\tilde{b}_g$ es el vector de medias de $\tilde{X}_{g_{norm}}$, $\tilde{\Sigma}_g$ es su matriz (diagonal) de desviaciones estándar, $\Delta b_g \coloneqq \tilde{b}_g - b_g$ es la disrupción en el vector de medias, y $\boldsymbol{x}_{p_{norm}} = \left(\boldsymbol{x}_p - \tilde{b}_g^T\right)\tilde{\Sigma}_g^{-1}$. Tanto $\tilde{b}_g$ como $\tilde{\Sigma}_g$ son desconocidos y se deben, por lo tanto, derivar expresiones para éstos también. El vector de medias distorsionado se calcula como $\tilde{b}_g = \frac{1}{\tilde{n}_g}\left(\tilde{X}_g\right)^T I_{\tilde{n}_g}$, lo que puede ser fácilmente convertido en:

$$\tilde{b}_g = \frac{n_g}{n_g + 1} b_g + \frac{1}{n_g + 1} \boldsymbol{x}_p^T \quad (34)$$

La ecuación (34) muestra que los pesos naturales, en el caso del vector de medias distorsionado, son $\frac{n_g}{n_g+1}$ y $\frac{1}{n_g+1}$ respectivamente para b_g y $\boldsymbol{x}_p$. Con el fin de obtener una expresión para la matriz de desviaciones estándar distorsionada, $\tilde{\Sigma}_g$, se lleva a cabo una sustracción (columna por columna) del vector de medias para $\tilde{X}_g$:

$$\tilde{X}_{g_{mean-centered}} = \tilde{X}_g - I_{\tilde{n}_g}\tilde{b}_g^T = \begin{bmatrix} X_g - I_{n_g}\tilde{b}_g^T \\ \boldsymbol{x}_p - \tilde{b}_g^T \end{bmatrix} \quad (35)$$

Adicionando y sustrayendo $I_{n_g}b_g^T$ al término $X_{g_{norm}}\Sigma_g - I_{n_g}\tilde{b}_g^T$ en la ecuación (35), ésta se puede rescribir así:

$$\tilde{X}_{g_{mean-centered}} = \begin{bmatrix} \left(X_g - I_{n_g}b_g^T\right) - I_{n_g}\Delta b_g^T \\ \boldsymbol{x}_p - \tilde{b}_g^T \end{bmatrix} \quad (36)$$

Sea:

$$\tilde{X}_{g_{mean-centered}}(:,i) = \begin{bmatrix} \left(X_g(:,i) - b_g(i)I_{n_g}\right) - \Delta b_g(i)I_{n_g} \\ \boldsymbol{x}_p(i) - \tilde{b}_g(i) \end{bmatrix} \quad (37)$$

la i-ésima columna de $\tilde{X}_{g_{mean-centered}}$, correspondiente a la i-ésima variable. Entonces la varianza de esta i-ésima variable será

$\tilde{\sigma}_{g_i}^2 = \frac{1}{\tilde{n}_g-1}\left(\tilde{X}_{g_{mean-centered}}(:,i)\right)^T \tilde{X}_{g_{mean-centered}}(:,i)$, lo cual puede ser escrito como:

$$\left(\tilde{n}_g - 1\right)\tilde{\sigma}_{g_i}^2$$
$$= \left[\left(X_g^T(:,i) - b_g(i)I_{n_g}^T\right) - \Delta b_g(i)I_{n_g}^T \quad \boldsymbol{x}_p(i) - \tilde{b}_g(i)\right]\begin{bmatrix} \left(X_g(:,i) - b_g(i)I_{n_g}\right) - \Delta b_g(i)I_{n_g} \\ \boldsymbol{x}_p(i) - \tilde{b}_g(i) \end{bmatrix} \quad (38)$$

Esto puede ser expandido como:

$$\left(\tilde{n}_g - 1\right)\tilde{\sigma}_{g_i}^2 = \left(X_g^T(:,i) - b_g(i)I_{n_g}^T\right)\left(X_g(:,i) - b_g(i)I_{n_g}\right) - \left(X_g^T(:,i) - b_g(i)I_{n_g}^T\right)\Delta b_g(i)I_{n_g} - \Delta b_g(i)I_{n_g}^T\left(X_g(:,i) - b_g(i)I_{n_g}\right) + \Delta b_c^2(i)I_{n_g}^T I_{n_g} + \left(\boldsymbol{x}_p(i) - \tilde{b}_g(i)\right)^2 \quad (39)$$

En la ecuación (39), nótese que $\left(X_g^T(:,i) - b_g(i)I_{n_g}^T\right)\left(X_g(:,i) - b_g(i)I_{n_g}\right) = \left(n_g - 1\right)\sigma_{g_i}^2$, $I_{n_g}^T I_{n_g} = n_g$, y que $\left(X_g^T(:,i) - b_g(i)I_{n_g}^T\right)\Delta b_g(i)I_{n_g} = \Delta b_g(i)I_{n_g}^T\left(X_g(:,i) - b_g(i)I_{n_g}\right)$. Entonces, la ecuación (39) se puede reducir a:

$$\left(\tilde{n}_g - 1\right)\tilde{\sigma}_{g_i}^2 = \left(n_g - 1\right)\sigma_{g_i}^2 - 2\Delta b_g(i)I_{n_g}^T\left(X_g(:,i) - b_g(i)I_{n_g}\right) + n_g\Delta b_g^2(i) + \left(\boldsymbol{x}_p(i) - \tilde{b}_g(i)\right)^2 \quad (40)$$

Además, recordando que $\tilde{n}_g = n_g + 1$, y notando que $I_{n_g}^T X_g(:,i) = I_{n_g}^T\left(b_g(i)I_{n_g}\right) = n_g b_g(i)$, el término $2\Delta b_g(i)I_{n_g}^T\left(X_g(:,i) - b_g(i)I_{n_g}\right)$ desaparece, obteniéndose finalmente:

$$\tilde{\sigma}_{g_i} = \sqrt{\frac{n_g - 1}{n_g}\sigma_{g_i}^2 + \Delta b_g^2(i) + \frac{1}{n_g}\left(\boldsymbol{x}_p(i) - \tilde{b}_g(i)\right)^2} \quad (41)$$

La ecuación (41) muestra que las varianzas (distorsionadas) de las variables en el grupo $\tilde{X}_g$ dependen de:

- Las varianzas originales en X_g, con un peso natural de $\frac{n_g - 1}{n_g}$.
- Los valores cuadrados de la nueva muestra (centrada en la media), $\left(\boldsymbol{x}_p(i) - \tilde{b}_g(i)\right)^2$, con un peso natural de $\frac{1}{n_g}$.
- Los valores cuadrados de la disrupción en el vector de medias, $\Delta b_g^2(i)$.

La ecuación (41) debe ser ejecutada para $i = 1,2, \dots m$, de manera que se construya:

$$\tilde{\Sigma}_g = \begin{bmatrix} \sigma_{g_1} & & \\ & \ddots & \\ & & \sigma_{g_m} \end{bmatrix} \quad (42)$$

Ahora que se tiene expresiones para $\tilde{\mathrm{b}}_g$ y $\tilde{\Sigma}_g$, se procede a buscar una expresión para la matriz de correlación distorsionada, la cual debería calcularse como $\tilde{S}_g = \frac{1}{\tilde{n}_g - 1} \tilde{X}_{g_{norm}}^T \tilde{X}_{g_{norm}}$. Combinando esto con la ecuación (33) se tiene:

$$\begin{aligned} &(\tilde{n}_g - 1)\tilde{S}_g \\ &= \begin{bmatrix} \tilde{\Sigma}_g^{-1} \left(\Sigma_g X_{g_{norm}}^T - \Delta b_g I_{n_g}^T \right) & \boldsymbol{x}_{p_{norm}}^T \end{bmatrix} \begin{bmatrix} \left(X_{g_{norm}} \Sigma_g - I_{n_g} \Delta b_g^T \right) \tilde{\Sigma}_g^{-1} \\ \boldsymbol{x}_{p_{norm}} \end{bmatrix} \quad (43) \end{aligned}$$

Expandiendo el producto en la ecuación (43) se tiene:

$$\begin{aligned} (\tilde{n}_g - 1)\tilde{S}_g = {} & \tilde{\Sigma}_g^{-1} \Sigma_g X_{g_{norm}}^T X_{g_{norm}} \Sigma_g \tilde{\Sigma}_g^{-1} - \tilde{\Sigma}_g^{-1} \Sigma_g X_{g_{norm}}^T I_{n_g} \Delta b_g^T \tilde{\Sigma}_g^{-1} \\ & - \tilde{\Sigma}_g^{-1} \Delta b_g I_{n_g}^T X_{g_{norm}} \Sigma_g \tilde{\Sigma}_g^{-1} + \tilde{\Sigma}_g^{-1} \Delta b_g I_{n_g}^T I_{n_g} \Delta b_g^T \tilde{\Sigma}_g^{-1} \\ & + \boldsymbol{x}_{p_{norm}}^T \boldsymbol{x}_{p_{norm}} \quad (44) \end{aligned}$$

Nótese que: $X_{g_{norm}}^T X_{g_{norm}} = (n_g - 1) S_g$, $\Sigma_g X_{g_{norm}}^T = X_g^T - b_g I_{n_g}^T$, $X_{g_{norm}} \Sigma_g = X_g - I_{n_g} b_g^T$ y $I_{n_g}^T I_{n_g} = n_g$. Por lo anterior, la ecuación (44) se puede rescribir como:

$$\begin{aligned} (\tilde{n}_g - 1)\tilde{S}_g = {} & (n_g - 1) \tilde{\Sigma}_g^{-1} \Sigma_g S_g \Sigma_g \tilde{\Sigma}_g^{-1} - \tilde{\Sigma}_g^{-1} \left(X_g^T - b_g I_{n_g}^T \right) I_{n_g} \Delta b_g^T \tilde{\Sigma}_g^{-1} \\ & - \tilde{\Sigma}_g^{-1} \Delta b_g I_{n_g}^T \left(X_g - I_{n_g} b_g^T \right) \tilde{\Sigma}_g^{-1} + n_g \tilde{\Sigma}_g^{-1} \Delta b_g \Delta b_g^T \tilde{\Sigma}_g^{-1} \\ & + \boldsymbol{x}_{p_{norm}}^T \boldsymbol{x}_{p_{norm}} \quad (45) \end{aligned}$$

Notando además que $X_g^T I_{n_g} = b_g I_{n_g}^T I_{n_g} = I_{n_g}^T X_g = I_{n_g}^T I_{n_g} b_g^T = n_g b_g$, el segundo y el tercer término en la ecuación (45) desaparecen, obteniéndose:

$$\tilde{S}_g = \frac{n_g - 1}{n_g}\tilde{\Sigma}_g^{-1}\Sigma_g S_g \Sigma_g \tilde{\Sigma}_g^{-1} + \tilde{\Sigma}_g^{-1}\Delta b_g \Delta b_g^T \tilde{\Sigma}_g^{-1} + \frac{1}{n_g}\boldsymbol{x}_{p_{norm}}^T \boldsymbol{x}_{p_{norm}} \quad (46)$$

La ecuación (46) muestra que la matriz de correlación distorsionada, $\tilde{S}_g$, depende de tres términos:

- $\tilde{\Sigma}_g^{-1}\Sigma_g S_g \Sigma_g \tilde{\Sigma}_g^{-1}$, que tiene en cuenta la contribución de la matriz de correlación no distorsionada S_g después de una actualización de desviación estándar, con un peso natural de $\frac{n_g - 1}{n_g}$.
- $\boldsymbol{x}_{p_{norm}}^T \boldsymbol{x}_{p_{norm}}$, que tiene en cuenta la contribución de la nueva muestra (normalizada) a la constitución de la matriz de correlación distorsionada, con un peso natural de $\frac{1}{n_g}$.
- $\tilde{\Sigma}_g^{-1}\Delta b_g \Delta b_g^T \tilde{\Sigma}_g^{-1}$, que considera el efecto de la disrupción de Σ_g y b_g en $\tilde{S}_g$.

La disrupción de la estructura de correlación se puede medir mediante la estimación de la desviación entre S_g y $\tilde{S}_g$, usando el estadístico $\varphi(S_g, \tilde{S}_g)$ (ver ecuación (13)), propuesto en la sección 3.2.2.

Como se ha dicho previamente, si la base de datos para el grupo g contiene un gran número de muestras, la integración de $\boldsymbol{x}_p$ apenas causará una pequeña disrupción en la estructura de correlación, aunque tenga características diferentes en comparación con las de las muestras contenidas en X_g. Por ejemplo, si X_g se compusiera de 200 muestras, el peso relativo natural de su vector de medias (b_c) para la constitución del vector de medias distorsionado ($\tilde{b}_g$) sería aproximadamente 0.995, mientras que el peso de la muestra sería (tan solo) aproximadamente 0.005.

Por otra parte, si los pesos fuesen calculados suponiendo que X_g se compone de unas pocas muestras, i.e., reemplazando n_g por n_g^{red} (tal que $n_g^{red} \ll n_g$) en los cocientes para calcular los pesos relativos, esos pesos resultan más equitativos, proveyendo

factores de ponderación para el cálculo de la matriz de correlación distorsionada, pero utilizando toda la información de las muestras originales de X_g, ya que ésta permanece contenida en b_g, Σ_g y S_g. Esto es equivalente a encontrar una base generatriz de pocas muestras/pacientes (n_g^{red}) que representa todas las características de X_g para luego incorporar $\boldsymbol{x}_p$ y evaluar la disrupción causada a la estructura de correlación, proporcionando una reducción dimensional artificial. Por ejemplo, si se calculan los pesos relativos como si X_g se compusiera de sólo tres muestras que exhiben todos los atributos del conjunto de datos original (es decir, haciendo $n_g^{red} = 3$), los pesos relativos para el cálculo del vector de medias distorsionado serían 0.75 y 0.25 respectivamente.

El umbral inferior para esta reducción dimensional artificial se encuentra al hacer $n_g^{red} = 2$ para el cálculo de los pesos relativos ponderados, ya que hacer $n_g^{red} = 1$ conllevaría a dejar de lado toda la información contenida en S_g para la estimación de $\tilde{S}_g$ (ver ecuación (46)), así como también pasa algo similar con las desviaciones estándar (ver ecuación (41)).

5.2. Escenario II: Integración de un Nuevo Conjunto de Muestras a una Base de Datos para la Determinación de la Distorsión causada a su Estructura de Correlación

Supóngase que se tiene ahora un conjunto $X_p^{\rho} \in \mathbb{R}^{n_p \times m}$ de n_p nuevas muestras (en lugar de una sola muestra, $\boldsymbol{x}_p^{\rho}$) que se quieren agregar a la base de datos. También debe hacerse el pretratamiento de datos para este conjunto de muestras, de manera que: $X_p = clr\left(c\left(BM\left(X_p^{\rho}\right)\right)\right)$. Utilizar las ecuaciones (34), (41) y (46) para evaluar la distorsión causada por la inclusión del nuevo conjunto X_p no sería práctico, ya que deberían emplearse n_p veces, i.e., una vez por cada muestra. En lugar de esto, deben

proporcionarse ecuaciones que permitan determinar la distorsión causada a la estructura de correlación debido a la integración de estas n_p nuevas muestras. Portnoy et al. (Portnoy *et al.*, 2016) desarrollaron ecuaciones para la actualización recursiva de (entre otros atributos estadísticos) los vectores de medias, matrices de desviación estándar y matrices de correlación ante la disponibilidad de un nuevo conjunto de muestras para incluir en la base de datos. Utilizando una notación congruente con la que se ha llevado a lo largo de este documento, dichas ecuaciones, con sus pesos naturales, se podrían resumir como sigue:

- Para el vector de medias distorsionado:

$$\tilde{b}_g = \frac{n_g}{n_g + n_p} b_g + \frac{n_p}{n_g + n_p} b_p \quad (47)$$

  donde $b_p = \frac{1}{n_p} \left(X_p\right)^T I_{n_p}$ es el vector de medias del nuevo conjunto de muestras y $I_{n_p} = [1 \quad 1 \quad \cdots \quad 1]^T \mathbb{R}^{n_p \times 1}$.

- Para la matriz de desviaciones estándar distorsionada:

$$\tilde{\Sigma}_g^2 = \frac{n_g - 1}{n_g + n_p - 1} \Sigma_g^2 + \frac{n_p - 1}{n_g + n_p - 1} \Sigma_p^2 + \frac{n_g}{n_g + n_p - 1} diag\left(\left(\Delta b_g \Delta b_g^T\right)(i,i)\right) \quad (48)$$

  donde Σ_p es la matriz de desviaciones estándar del nuevo conjunto de muestras, $\Delta b_g = \tilde{b}_g - b_g$, y $diag\left(\left(\Delta b_g \Delta b_g^T\right)(i,i)\right)$ es una matriz diagonal que contiene únicamente los elementos diagonales de $\Delta b_g \Delta b_g^T$.

- Para la matriz de correlaciones:

$$\tilde{S}_g = \frac{n_g - 1}{n_g + n_p - 1} \tilde{\Sigma}_g^{-1} \Sigma_g S_g \Sigma_g \tilde{\Sigma}_g^{-1} + \frac{n_p - 1}{n_g + n_p - 1} S_p + \frac{n_g}{n_g + n_p - 1} \tilde{\Sigma}_g^{-1} \Delta b_g \Delta b_g^T \tilde{\Sigma}_g^{-1} \quad (49)$$

  donde $S_p = \frac{1}{n_p - 1} \left(X_p^T - b_p I_{n_p}^T\right)^T \left(X_g^T - b_g I_{n_g}^T\right)$.

De nuevo, si se quiere hacer una reducción dimensional artificial del número de muestras contenidas en la base de datos X_g, con el fin de darle una mayor ponderación a las nuevas muestras, basta con reemplazar n_g por n_g^{red} (tal que $n_g^{red} \ll n_g$) en los cocientes para calcular los pesos relativos en las ecuaciones (47), (48) y (49).

En futuras investigaciones, se abordará el uso de esta técnica de reducción dimensional, junto con el estadístico propuesta en la sección 3.2.2., para crear una técnica de clasificación de nuevas muestras.

6. INTERGRACIÓN DE LAS TÉCNICAS DESARROLLADAS

Este capítulo explica cómo se hace la integración de las técnicas diseñadas en una plataforma computacional con el fin de crear una herramienta para el análisis de datos provenientes de experimentos de secuenciación o para aplicaciones industriales. Se define también la secuencia lógica de operación de esta herramienta, así como sus resultados y gráficos generados para ilustrar estos resultados para el usuario. La implementación se realiza en R.

La Figura 10 muestra una representación esquemática, en un diagrama de flujo, de la integración de las técnicas propuestas. El usuario debe proporcionar los dos conjuntos de datos sin pretratamiento: el conjunto de controles, $X_c^{\rho} \in \mathbb{R}^{n_c \times m}$, y el conjunto de casos, $X_v^{\rho} \in \mathbb{R}^{n_v \times m}$, concatenados en: $X_T^{\rho} = \begin{bmatrix} X_c^{\rho} \\ X_v^{\rho} \end{bmatrix} \in \mathbb{R}^{(n_c+n_v) \times m}$. Además, se le pregunta al usuario si los datos son o no de naturaleza composicional. En caso afirmativo, se ejecuta todo el pretratamiento de datos explicado en detalle en la sección 3.2.1. De lo contrario, sólo se lleva a cabo un auto-escalamiento para ambos grupos, como sigue:

$$X_{g_{as}} = \left(X_g - I_{n_g} b_g^T\right) \Sigma_g^{-1} \quad (50)$$

donde $b_g = \frac{1}{n_g} \left(X_g\right)^T I_{n_g}$ es un vector columna que contiene las medias de todas las m variables para el grupo g (controles o casos), $I_{n_g} = [1\ 1\ \ldots 1]^T \in \mathbb{R}^{n_g}$, y Σ_g es una matriz diagonal que contiene las desviaciones estándar de las m variables para el grupo g.

Luego del pretratamiento o el auto-escalamiento, según sea el caso, se ejecuta la evaluación de similitud de estructuras de correlación usando la técnica propuesta, la cual lleva embebida la técnica de reducción dimensional que se desee usar.

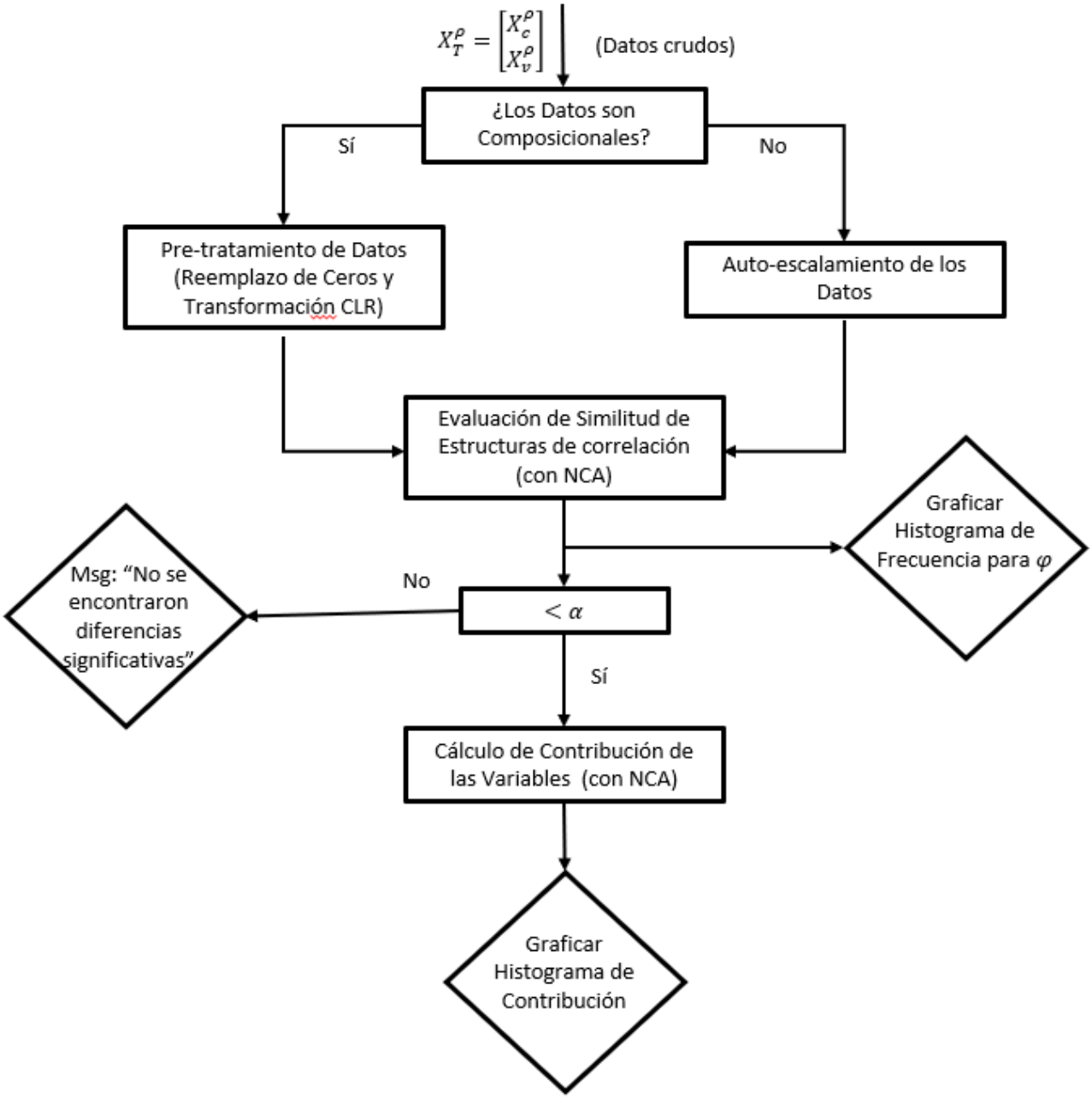

Figura 10 – Diagrama de flujo esquematizando la integración de las técnicas propuestas en una plataforma computacional para la constitución de un pipeline de análisis de datos provenientes de experimentos de secuenciación.

La implementación en R debe mostrar el histograma de frecuencia construido a partir de la ejecución del bootstrapping junto con el valor del estadístico de prueba calculado con los conjuntos X_c^{ρ} y X_v^{ρ}. Además, la implementación debe mostrar el valor p obtenido a partir de la prueba de hipótesis no paramétrica propuesta. Luego este valor p debe compararse con la probabilidad nominal de error tipo I, α, la cual es usualmente fijada en 0.05 pero puede ser cambiada por el usuario. Si el valor p resulta ser mayor o igual que α, se imprime un aviso para el usuario indicándole que no se encontraron diferencias significativas entre las estructuras de correlación de los conjuntos de datos evaluados. De lo contrario, se le indica que se han encontrado diferencias significativas y se procede con el cálculo de las contribuciones de cada variable a las diferencias, el cual se hace con la técnica de análisis diferencial propuesta en el capítulo 4.

Luego del análisis diferencial, se debe generar una tabla que contenga las contribuciones individuales de cada variable. Además, se genera un histograma de contribución mostrando las variables que más contribuyeron. Por defecto se muestran las 10 variables más contribuyentes, pero el usuario debe poder cambiar este número si desea ver más o menos variables en el histograma.

Debido a la naturaleza secuencial de los análisis que se hacen con las técnicas propuestas, la integración de éstas resulta muy simple, como se puede apreciar en la Figura 10. A este Pipeline se le ha dado el acrónimo de NCA, que abrevia: Non-parametric compositional-data assessment.

7. VALIDACIÓN DE TÉCNICAS DESARROLLADAS CON DATOS REALES DE SECUENCIACIÓN

En este capítulo, se analizarán dos casos estudio con datos provenientes de experimentos de secuenciación de 16S rRNA con el fin de demostrar la efectividad de las técnicas desarrolladas. En ambos casos estudio, se excluyeron de los análisis las OTUs que no estaban presentes en al menos el 50% de las muestras, con el fin de reducir la tasa de falsos positivos, tal como lo hicieron (Buendía *et al.*, 2018; San-Juan-Vergara *et al.*, 2018).

7.1. Caso Estudio I: Comunidades Bacterianas en Lagos Pantanosos

Linz et al. (Linz *et al.*, 2017) condujeron un estudio sobre las comunidades bacterianas presentes en ocho lagos pantanosos. Meticulosamente, Linz et al. recopilaron muestras composicionales de la microbiota y las clasificaron dependiendo de la locación (i.e., el lago al que pertenecen), estratificación del agua (i.e., capa) y estación del año. Para propósitos ilustrativos, con el fin de demostrar la funcionalidad de las técnicas propuestas, sólo se evalúan aquí los cambios en las estructuras de correlación de la comunidad bacteriana entre (a) dos capas del mismo lago y entre (b) dos lagos considerando la misma capa. En particular, se comparará (a) la capa Epilimnion del lago Trout Bog (TBE) con la capa Hipolimnion del mismo lago (TBH), y la capa Epilimnion de TBE con la misma capa del lago South Sparkling (SSE). Los datos usados están disponibles en el paquete de R OTUtable (Linz *et al.*, 2017).

Las Figuras 11a, 11b y 11c muestran mapas de calor representando gráficamente las estructuras de correlación de los conjuntos de datos de TBE, TBH y SSE respectivamente. Se encontró que la técnica propuesta y la de Krzanowski detectaron cambios significativos en las estructuras de correlación de las comunidades bacterianas para los casos analizados, como se resume en la Tabla 3.

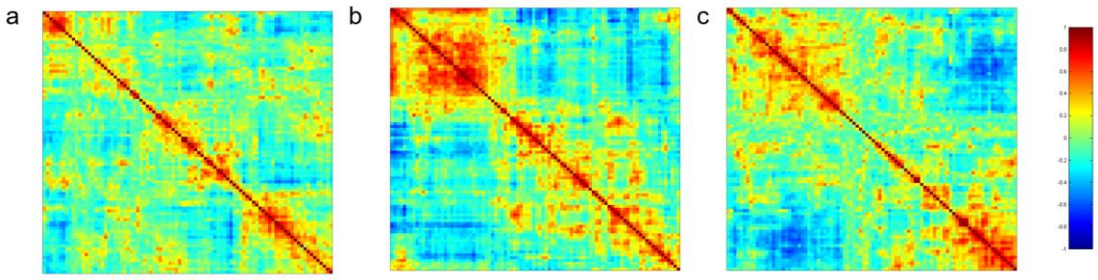

Figura 11 - Mapas de calor de correlación en caso estudio I para (a) TBE; (b) TBH; (c) SSE.

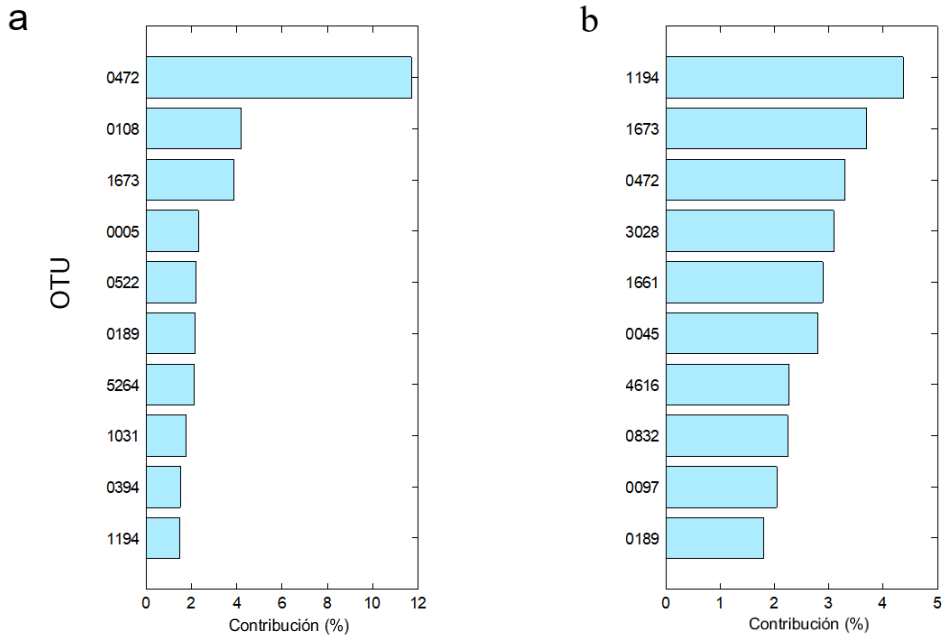

Figura 12 - Top 10 de las variables que más contribuyen a la diferenciación en caso estudio I para (a) TBE vs TBH y (b) TBE vs SSE.

Las Figuras 12a y 12b muestran la lista de las 10 variables que más contribuyen a la diferenciación de las estructuras de correlación para los dos escenarios evaluados en este caso estudio.

Tabla 3 – Resultados de la evaluación de similitud en estructuras de correlación, caso estudio I.

Escenario	*Valor p*		
	Krzanowski	**dna**	**NCA**
TBE vs TBH	0.0000	0.4703	0.0072
TBE vs SSE	0.0033	0.7953	0.0003

Tabla 4 -Taxonomía de los OTUs relevantes, Caso Estudio I.

Código	Taxonomía						
	Reino	**Filo**	**Clase**	**Orden**	**Lineaje**	**Clado**	**Tribu**
Otu0005	k__Bacteria(100)	p__Proteobacteria(100)	c__Betaproteobacteria(100)	o__Burkholderiales(100)	betI(95)	unclassified	unclassified
Otu0045	k__Bacteria(100)	p__Proteobacteria(100)	c__Gammaproteobacteria(100)	o__Methylococcales(100)	gamI(100)	unclassified	unclassified
Otu0097	k__Bacteria(100)	p__Proteobacteria(100)	c__Betaproteobacteria(100)	o__Burkholderiales(100)	betII(100)	Pnec(100)	PnecC(100)
Otu0108	k__Bacteria(100)	p__Proteobacteria(100)	c__Betaproteobacteria(100)	o__Burkholderiales(100)	betI(98)	betI-A(95)	Lhab-A1(80)
Otu0189	k__Bacteria(100)	p__Actinobacteria(100)	c__Actinobacteria(100)	o__Actinomycetales(100)	acI(100)	acI-B(100)	acI-B2(100)
Otu0394	k__Bacteria(100)	p__Proteobacteria(96)	c__Gammaproteobacteria(88)	o__Xanthomonadales(70)	f__Sinobacteraceae(70)	unclassified	unclassified
Otu0472	k__Bacteria(100)	p__Proteobacteria(100)	c__Betaproteobacteria(100)	o__Burkholderiales(100)	betI(98)	betI-A(98)	Lhab-A4(81)
Otu0522	k__Bacteria(100)	p__Proteobacteria(100)	c__Alphaproteobacteria(100)	o__Rhizobiales(100)	alfI(100)	alfI-A(100)	alfI-A1(100)
Otu0832	k__Bacteria(100)	p__Verrucomicrobia(100)	c__Opitutae(100)	o__Opitutales(100)	f__Opitutaceae(100)	unclassified	unclassified
Otu1031	k__Bacteria(100)	p__Proteobacteria(100)	c__Betaproteobacteria(100)	o__Methylophilales(99)	betIV(99)	unclassified	unclassified
Otu1194	k__Bacteria(100)	p__Proteobacteria(100)	c__Betaproteobacteria(92)	o__Methylophilales(81)	unclassified	unclassified	unclassified
Otu1661	k__Bacteria(100)	p__Proteobacteria(100)	c__Deltaproteobacteria(100)	o__Desulfobacterales(100)	f__Desulfobulbaceae(99)	unclassified	unclassified
Otu1673	k__Bacteria(100)	p__Actinobacteria(100)	c__Acidimicrobiia(100)	o__Acidimicrobiales(100)	acV(100)	acV-A(100)	acV-A2(100)
Otu3028	k__Bacteria(100)	p__Verrucomicrobia(100)	c__Opitutae(100)	o__Opitutales(100)	f__Opitutaceae(100)	unclassified	unclassified
Otu4616	k__Bacteria(100)	p__Bacteroidetes(100)	c__[Saprospirae](100)	o__[Saprospirales](100)	f__Chitinophagaceae(100)	unclassified	unclassified
Otu5264	k__Bacteria(100)	p__Actinobacteria(97)	c__Acidimicrobiia(96)	o__Acidimicrobiales(96)	unclassified	unclassified	unclassified

Como se observa, NCA detectó cambios significativos en las estructuras de correlación de la comunidad bacteriana para ambos escenarios (TBE vs TBH y TBE vs SSE) (ver Tabla 3). Por el contrario, las pruebas de Krzanowski y dna no encontraron diferencias significativas entre las estructuras de correlación en ninguno de los escenarios. Además, el algoritmo NCA pudo identificar las 10 principales variables contribuyentes para ambos escenarios de comparación, como se muestra en las Figuras 6d y 6e. En estas Figuras, los nombres de los OTUs están codificados. Se muestran los nombres reales en la Tabla 4.

7.2. Caso Estudio II: Microbiota Intestinal en Pacientes Asmáticos de una Región Tropical

Para este caso estudio se utilizan las matrices de conteos de OTUs obtenidas por Buendía et al. (Buendía *et al.*, 2018) en un estudio reciente sobre la relación entre el asma y la microbiota intestinal en pacientes de la región caribe colombiana. Los autores clasificaron a los pacientes en tres categorías basados en la severidad del asma: Sin obstrucción en las vías aéreas (NAO, por sus siglas en inglés), obstrucción reversible de las vías aéreas (RAO) y obstrucción fija en las vías aéreas (FAO). La Figura 13 muestra los mapas de calor de correlación para estos tres grupos,

Para ir en línea con el estudio de Buendía et al., se compararán: el grupo FAO vs NAO y FAO vs RAO, ya que el grupo FAO constituye un conjunto de datos de control. La similitud de estructuras de correlación, para ambos escenarios de comparación, es evaluada usando las técnicas de Krzanowski, dna y NCA (propuesta). Los resultados esta evaluación están resumidos en la Tabla 5.

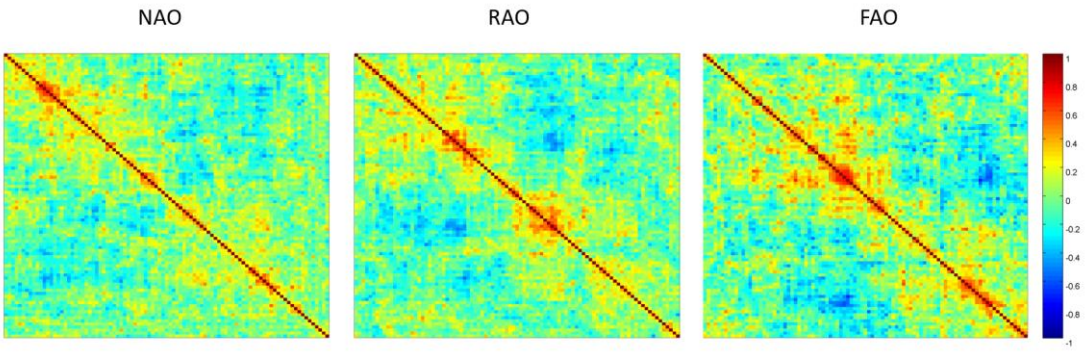

Figura 13 - Mapas de calor de correlación en caso estudio II para NAO, RAO y FAO.

Tabla 5 – Resultados de la evaluación de similitud en estructuras de correlación, caso estudio II.

Escenario	Valor *p*		
	Krzanowski	**dna**	**NCA**
FAO vs NAO	0.5424	0.8102	0.0701
FAO vs RAO	0.9256	0.2428	0.5437

Al comparar los grupos FAO y NAO, en contraste con la técnica dna (valor $p = 0.8040$) y la de Krzanowski (valor $p = 0.5424$), NCA estuvo muy cerca de dilucidar un cambio significativo en la estructura de correlación global (valor $p = 0.0701$) usando una probabilidad de error tipo I del 5%, mientras que para una taza de error tipo I (más laxa) del 10% se habría detectado un cambio significativo. La inhabilidad de todos los métodos para detectar diferencias significativas podría ser explicada por el limitado número de pacientes 42, 66 y 74 para los grupos FAO, NAO y RAO, respectivamente) comparado con el número de variables bajo estudio (93 luego de eliminar aquellos OTUs con presencia en menos del 50% de las muestras). Además, es notable, en las Figuras 6a, 6b, 6c y 6d, que ésta es la región en la que NCA mostró el desempeño más bajo en los experimentos con datos sintéticos. A pesar de lo anteriormente expuesto, el

método propuesto (NCA) tiende a la detección de diferencias de una manera que no tiene comparación con los otros métodos usados.

Por otra parte, la comparación entre los grupos FAO y RAO no reveló ninguna diferencia significativa sin importar el método usado (Tabla 5). Este resultado está alineado con el reporte original (Buendía *et al.*, 2018), y puede ser explicado por el hecho de que las condiciones de los grupos FAO y RAO son más similares entre sí que aquellas para los grupos FAO y NAO. La Figura 14 muestra el top 10 de las variables más contribuyentes detectadas por el método propuesto para el caso FAO vs NAO.

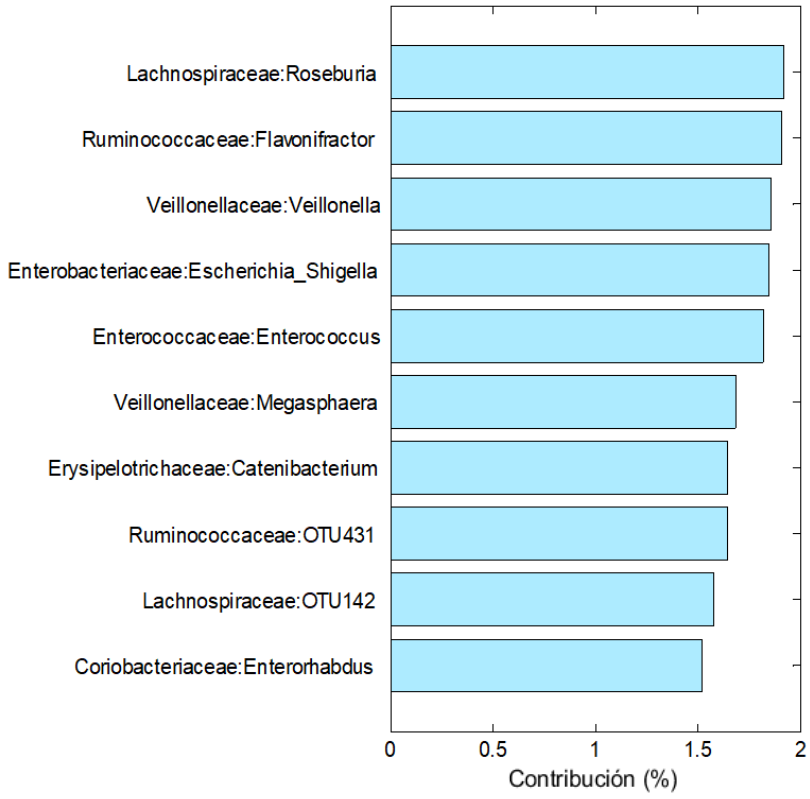

Figura 14 - Top 10 de las variables que más contribuyen a la diferenciación en caso estudio II para FAO vs NAO.

Es interesante notar que cinco OTUs en esta lista también están incluidas en el top 10 de las variables diferencialmente abundantes, reportadas en el estudio original. Esto indica que estas OTUs no solamente cambian significativamente su abundancia relativa en las comunidades bacterianas de la microbiota intestinal de los pacientes, sino que también cambian su interacción con los otros OTUs cuando se cambian las condiciones biológicas de FAO a NAO.

8. APLICACIÓN DE LAS TÉCNICAS PROPUESTAS CON DATOS DE PROCESOS INDUSTRIALES

Los datos de naturaleza composicional se pueden encontrar en distintas ramas de la ciencia. En química, por ejemplo, frecuentemente se busca determinar la composición de una sustancia en términos de las proporciones relativas de cada elemento que compone dicha sustancia, sin importar la cantidad (masa) total de la muestra que se analiza (Buccianti and Pawlowsky-Glahn, 2005). Análisis similares son llevados a cabo en geología (Bacon-Shone, 2011) para la determinación de la composición química de rocas y minerales, en arqueología para caracterizar la composición química de sitios, excavaciones y objetos (Baxter and Freestone, 2006; Miriello *et al.*, 2010), en la industria de alimentos y bebidas (Lipp and Anklam, 1998; Francis and Newton, 2005) para determinar la composición nutricional de los productos, en ciencias del medioambiente (Filzmoser, Hron and Reimann, 2009), economía (Fry, Fry and McLaren, 2000), entre otros campos.

La amplia variedad de ramas de la ciencia en las que se pueden usar técnicas de análisis de datos composicionales hace de los métodos desarrollados en este documento herramientas versátiles, cuyo uso en campos distintos a las ciencias biomédicas es factible y natural.

Por otra parte, si se prescindiese de usar (en la etapa de preprocesamiento) el algoritmo de Bayesiano de reemplazo de ceros y la transformación CLR de Aitchison, las técnicas desarrolladas se podrían usar en el campo de la detección, identificación y diagnóstico de fallas (Chiang, Russell and Braatz, 2000; Russell, Chiang and Braatz, 2012), así como también en aplicaciones de machine learning para la caracterización de variables predictoras para su posterior uso en modelos, clasificadores, etc.

En esta sección se presentará un tercer caso estudio que busca ilustrar el uso de las técnicas propuestas en una aplicación industrial: Detección de Cambio de Condición de Operación en una Red de Transporte de Gas Natural en la Costa Norte Colombiana.

8.2. Caso Estudio III: Detección de Cambio de Condición de Operación en una Red de Transporte de Gas Natural de PROMIGAS S.A. E.S.P. en la Costa Norte de Colombia

Para llevar a cabo este caso estudio, se extrae un conjunto de datos que corresponden a las señales de los sensores del primer segmento del sistema de transmisión de la empresa PROMIGAS S.A. E.S.P. (Ballena – La Mami) durante un período de 31 días (marzo de 2018). Este segmento se ilustra en la Figura 15, que fue tomada del trabajo de Pinzón et al. (Pinzón *et al.*, 2018).

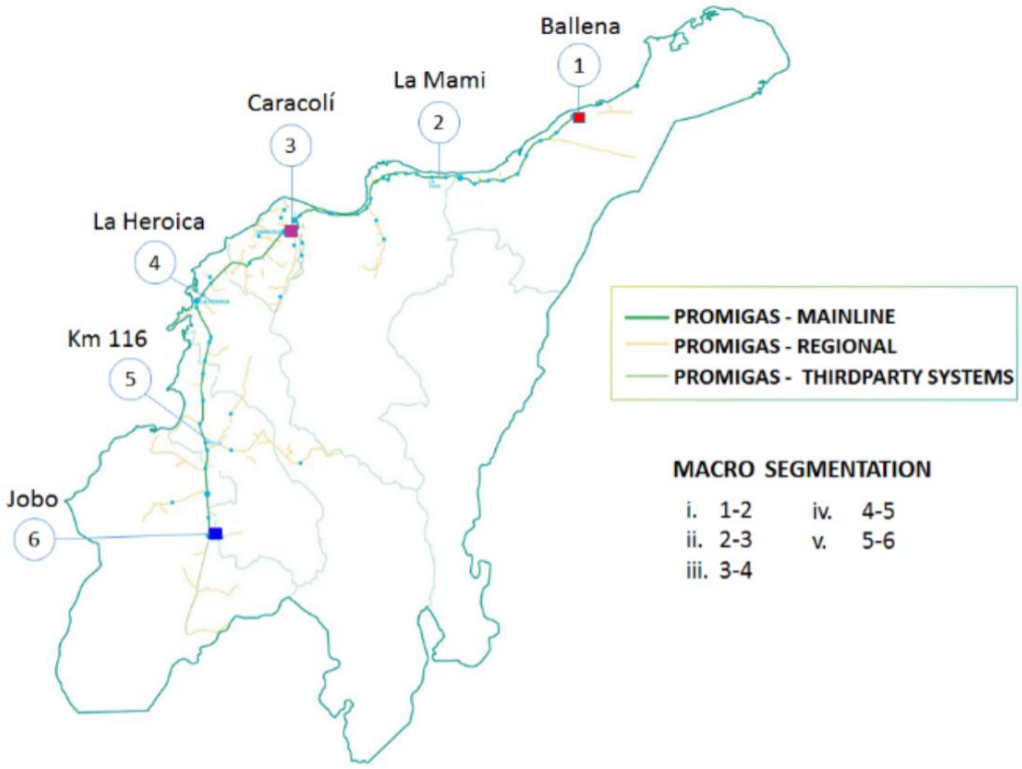

Figura 15 - Ilustración del sistema de transporte de gas natural de PROMIGAS S.A. E.S.P. (Pinzón et al., 2018).

Pinzón et al. (Pinzón *et al.*, 2018) reportaron la existencia de dos condiciones normales de operación (NOCs, por sus siglas en Inglés) distintas. Para este caso estudio se toman los conjuntos de datos correspondientes a NOC1 y NOC2 como X_c y X_v,

respectivamente. Se ejecuta NCA sin usar el algoritmo de Bayesiano de reemplazo de ceros ni la transformación CLR de Aitchison, ya que los datos no son de naturaleza composicional. El histograma de frecuencias para el estadístico φ se muestra en la Figura 16.

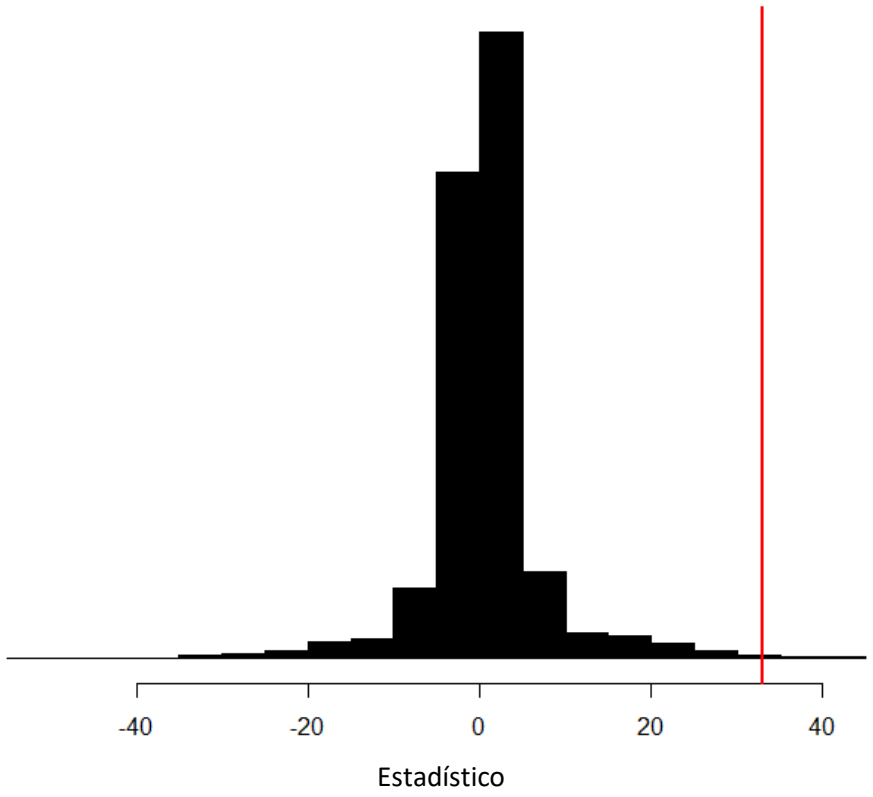

Figura 16 - Histograma de frecuencias de estadístico φ para Caso Estudio III (valor p = 0.004).

NCA es capaz de identificar una diferencia significativa, con un valor p de 0.004, entre las estructuras de correlación de X_c y X_v (correspondientes a NOC1 y NOC2, respectivamente).

La Figura 17 muestra las 5 variables, identificadas por NCA, que más contribuyen a las diferencias estructurales encontradas entre NOC1 y NOC2.

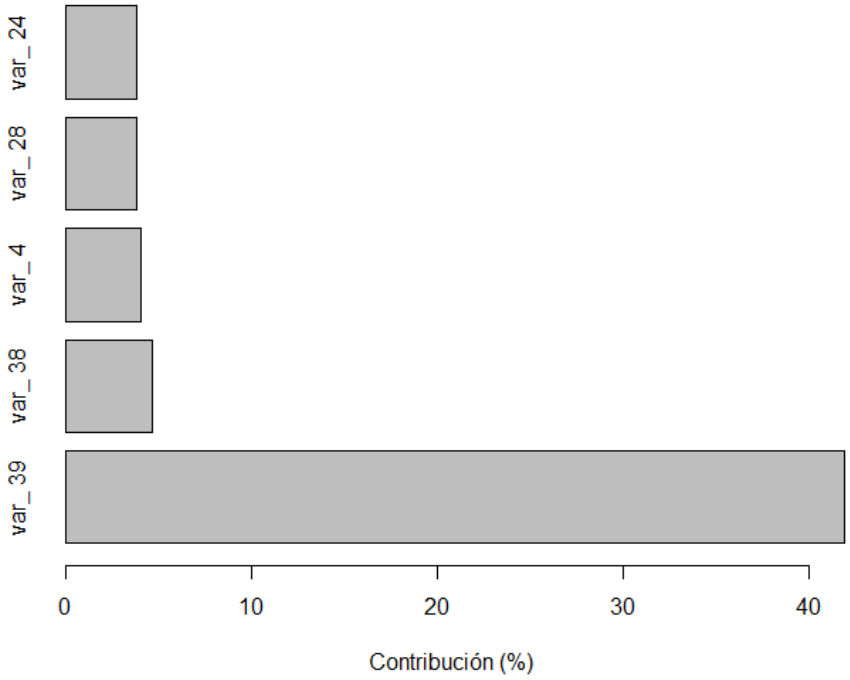

Figura 17 - Top 5 de las variables que más contribuyen a la diferenciación en Caso Estudio III entre NOC1 y NOC2.

En la Figura 17, las variables tienen nombres codificados, ya que los nombres que tienen en la base de datos de PROMIGAS son extensos, lo que dificulta su ilustración. En la Tabla 6 se muestran los nombres no codificados es estas variables.

Tabla 6 - Codificación de variables Caso Estudio III.

Nombre Codificado	Variable (Nombre en Base de Datos de PROMIGAS)
var_39	PALOM_FLOWRT_TOTAL
var_38	PALOMINO_V_PRES_L24A
var_4	BAL2_TEMPER_LIVE
var_28	MINGUEO20_V_PRES_L20A
var_24	DIBULLA_V_PRES_L20E

Los resultados de este caso estudio demuestran la aplicabilidad de las técnicas desarrolladas en este trabajo en aplicaciones industriales.

9. CONCLUSIONES

Un sistema es un grupo de entidades que interactúan o se interrelacionan para formar un todo. Un sistema es delineado por sus fronteras espaciales y temporales, es rodeado e influenciado por su ambiente, es declarado en su funcionamiento, y es descrito por su propósito y estructura.

La habilidad de comparar y entender las diferencias en la estructura de dos sistemas caracterizados por un gran número de variables es relevante para entender fenómenos complejos, particularmente en fenómenos de alta dimensionalidad como sistemas biológicos y clínicos. Cuando los datos que describen un grupo de entidades son composicionales, surgen limitaciones y desafíos adicionales con métodos tradicionales basados en la correlación de Pearson. El método no paramétrico de evaluación de datos composicionales (NCA, por sus siglas en inglés) propuesto en este documento, permite superar estos desafíos y demuestra un rendimiento superior a otros dos métodos (disponibles en la literatura) en la mayoría de los escenarios evaluados.

El NCA permite evaluar la similitud de la estructura global de correlación para redes biológicas, calcular la contribución de cada variable individual (por ejemplo, genes u OTUs) a los cambios encontrados, y señalar aquellas variables que más han cambiado sus interacciones de red cuando dos conjuntos de muestras se comparan. Aunque NCA se ha desarrollado para analizar conjuntos de datos biológicos provenientes de experimentos de secuenciación, su extensión a otros tipos de datos de naturaleza no composicional y no necesariamente provenientes de experimentos biológicos es posible y sencilla.

Basándose en simulación estadística, se condujeron dos experimentos para evaluar el desempeño de NCA, y compararlo con el de otras dos técnicas del estado del arte, en términos de la tasa de error tipo I y la potencia estadística. En general, los resultados muestran que NCA se desempeña mejor que los otros métodos evaluados para amplia región de dimensiones. El desempeño más bajo de la técnica se da cuando el número de muestras es muy pequeño. Por otra parte, se ejemplificó, con datos reales de

experimentos de secuenciación, la funcionalidad de NCA y se demostró su efectividad. Los resultados obtenidos con NCA se alinean con lo reportado por los autores de los reportes originales (Linz *et al.*, 2017; Buendía *et al.*, 2018).

Se diseñó además una técnica para la inclusión y ponderación, a través de una reducción dimensional artificial, nuevas muestras a una base de datos de conteos de OTUs previamente procesada. En futuras investigaciones, se abordará el uso de la técnica de reducción dimensional, propuesta en el capítulo 5, junto con el estadístico propuesto en la sección 3.2.2, para crear una técnica de clasificación de nuevas muestras en los grupos de control o casos (c o v, respectivamente), considerando la disrupción causada por la inclusión de éstas a la estructura de correlación de cada grupo.

Las técnicas propuestas en este manuscrito han sido concebidas para el análisis de datos de origen biológico. Sin embargo, su extensión y aplicación en otras ramas de la ciencia es factible, tal como se ha discutido en Capítulo 8.

REFERENCIAS

Aitchison, J. (1982) 'The statistical analysis of compositional data', *Journal of the Royal Statistical Society. Series B (Methodological)*. JSTOR, pp. 139–177.

Amanian, K. *et al.* (2007) 'Soft Sensor Based on Dynamic Principal Component Analysis and Radial Basis Function Neural Network for Distillation Column', in *Proceedings of the World Congress on Engineering and Computer Science*.

Amatya, A. and Demirtas, H. (2017) 'PoisNor: An R package for generation of multivariate data with Poisson and normal marginals', *Communications in Statistics - Simulation and Computation*. Taylor & Francis, 46(3), pp. 2241–2253. doi: 10.1080/03610918.2015.1039854.

Anders, S. and Huber, W. (2010) 'Differential expression analysis for sequence count data', *Genome biology*. BioMed Central, 11(10), p. R106.

Ardui, S. *et al.* (2018) 'Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics', *Nucleic acids research*. Oxford University Press, 46(5), pp. 2159–2168.

Bacon-Shone, J. (2011) 'A short history of compositional data analysis', *Compositional Data Analysis*. Wiley Online Library, pp. 1–11.

Baxter, M. J. and Freestone, I. C. (2006) 'Log-ratio compositional data analysis in archaeometry', *Archaeometry*. Wiley Online Library, 48(3), pp. 511–531.

Belilovsky, E., Varoquaux, G. and Blaschko, M. B. (2016) 'Testing for differences in gaussian graphical models: applications to brain connectivity', in *Advances in Neural Information Processing Systems*, pp. 595–603.

Bethge, M. and Berens, P. (2008) 'Near-maximum entropy models for binary neural representations of natural images', in *Advances in neural information processing systems*, pp. 97–104.

Buccianti, A. and Pawlowsky-Glahn, V. (2005) 'New perspectives on water chemistry and compositional data analysis', *Mathematical Geology*. Springer, 37(7), pp. 703–727.

Buendía, E. *et al.* (2018) 'Gut microbiota components are associated with fixed airway obstruction in asthmatic patients living in the tropics', *Scientific reports*. Nature Publishing Group, 8(1), p. 9582.

Buermans, H. P. J. and Den Dunnen, J. T. (2014) 'Next generation sequencing technology: advances and applications', *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*. Elsevier, 1842(10), pp. 1932–1941.

Carin, L. *et al.* (2012) 'High-dimensional longitudinal genomic data: an analysis used for monitoring viral infections', *IEEE signal processing magazine*. IEEE, 29(1), pp. 108–123.

Chiang, L. H., Russell, E. L. and Braatz, R. D. (2000) 'Fault diagnosis in chemical processes using Fisher discriminant analysis, discriminant partial least squares, and principal component analysis', *Chemometrics and intelligent laboratory systems*. Elsevier, 50(2), pp. 243–252.

Choi, S. W. and Lee, I.-B. (2004) 'Nonlinear dynamic process monitoring based on dynamic kernel PCA', *Chemical engineering science*. Elsevier, 59(24), pp. 5897–5908.

Chung, D. and Keles, S. (2010) 'Sparse partial least squares classification for high dimensional data', *Statistical applications in genetics and molecular biology*, 9(1).

Class, C. A. *et al.* (2018) 'iDINGO - integrative differential network analysis in genomics with Shiny application', *Bioinformatics*. Oxford University Press, 34(7), pp. 1243–1245.

Cole, N. (1968) 'The likelihood ratio test of the equality of correlation matrices', *The LL Thurstone Psychometric Laboratory*. University of North Carolina.

Collins, F. S. and Mansoura, M. K. (2001) 'The human genome project', *Cancer*, 91(s 1), pp. 221–225.

Van Dongen, S., Abreu-Goodger, C. and Enright, A. J. (2008) 'Detecting microRNA binding and siRNA off-target effects from expression data', *Nature methods*. Nature Publishing Group, 5(12), p. 1023.

Efron, B. and Tibshirani, R. (1986) 'Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy', *Statistical science*. JSTOR, pp. 54–75.

Erb, I. and Notredame, C. (2016) 'How should we measure proportionality on relative gene expression data?', *Theory in Biosciences*. Springer, 135(1–2), pp. 21–36.

Filzmoser, P., Hron, K. and Reimann, C. (2009) 'Univariate statistical analysis of environmental (compositional) data: problems and possibilities', *Science of the Total Environment*. Elsevier, 407(23), pp. 6100–6108.

Francis, I. L. and Newton, J. L. (2005) 'Determining wine aroma from compositional data', *Australian Journal of Grape and Wine Research*. Wiley Online Library, 11(2), pp. 114–126.

Fry, J. M., Fry, T. R. L. and McLaren, K. R. (2000) 'Compositional data analysis and zeros in micro data', *Applied Economics*. Taylor & Francis, 32(8), pp. 953–959.

Gill, R., Datta, Somnath and Datta, Susmita (2010) 'A statistical framework for differential network analysis from microarray data', *BMC bioinformatics*. BioMed Central, 11(1), p. 95.

Gill, R., Datta, Somnath and Datta, Susmita (2014) 'dna: An R package for differential network analysis', *Bioinformation*. Biomedical Informatics Publishing Group, 10(4), p. 233.

Goodwin, S., McPherson, J. D. and McCombie, W. R. (2016) 'Coming of age: ten years of next-generation sequencing technologies', *Nature Reviews Genetics*. Nature Publishing Group, 17(6), p. 333.

Herault, J. and Jutten, C. (1986) 'Space or time adaptive signal processing by neural network models', in *AIP conference proceedings*, pp. 206–211.

Herd, S. A. *et al.* (2014) 'A neural network model of individual differences in task switching

abilities', *Neuropsychologia*. Elsevier, 62, pp. 375–389.

Hood, L. E., Hunkapiller, M. W. and Smith, L. M. (1987) 'Automated DNA sequencing and analysis of the human genome', *Genomics*. Elsevier, 1(3), pp. 201–212.

Janková, J. and van de Geer, S. (2017) 'Honest confidence regions and optimality in high-dimensional precision matrix estimation', *Test*. Springer, 26(1), pp. 143–162.

Jeng, J.-C. (2010) 'Adaptive process monitoring using efficient recursive PCA and moving window PCA algorithms', *Journal of the Taiwan Institute of Chemical Engineers*. Elsevier, 41(4), pp. 475–481.

Jennrich, R. I. (1970) 'An asymptotic χ2 test for the equality of two correlation matrices', *Journal of the American Statistical Association*. Taylor & Francis, 65(330), pp. 904–912.

Jiang, B. and Braatz, R. D. (2017) 'Fault detection of process correlation structure using canonical variate analysis-based correlation features', *Journal of Process Control*. Elsevier, 58, pp. 131–138.

Josserand, T. M. (2008) 'Classification of gene expression data using PCA-based fault detection and identification', in *Genomic Signal Processing and Statistics, 2008. GENSiPS 2008. IEEE International Workshop on*, pp. 1–4.

Juric, D. *et al.* (2007) 'Differential gene expression patterns and interaction networks in BCR-ABL--positive and--negative adult acute lymphoblastic leukemias', *Journal of clinical oncology*. American Society of Clinical Oncology, 25(11), pp. 1341–1349.

Kim, H., Golub, G. H. and Park, H. (2006) 'Missing value estimation for DNA microarray gene expression data: local least squares imputation', *Bioinformatics*. Oxford University Press, 22(11), pp. 1410–1411.

Kohavi, R. and others (1995) 'A study of cross-validation and bootstrap for accuracy estimation and model selection', in *Ijcai*, pp. 1137–1145.

Krzanowski, W. J. (1993) 'Permutational tests for correlation matrices', *Statistics and Computing*. Springer, 3(1), pp. 37–44.

Kullback, S. (1967) 'On testing correlation matrices', *Applied Statistics*. JSTOR, pp. 80–85.

Kurtz, Z. D. *et al.* (2015) 'Sparse and compositionally robust inference of microbial ecological networks', *PLoS computational biology*. Public Library of Science, 11(5), p. e1004226.

Langmead, B., Hansen, K. D. and Leek, J. T. (2010) 'Cloud-scale RNA-sequencing differential expression analysis with Myrna', *Genome biology*. BioMed Central, 11(8), p. R83.

Larntz, K. and Perlman, M. D. (1985) 'A simple test for the equality of correlation matrices', *Rapport technique, Department of Statistics, University of Washington*, 141.

Li, J. and Ji, L. (2005) 'Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix', *Heredity*. Nature Publishing Group, 95(3), pp. 221–227.

Linz, A. M. *et al.* (2017) 'Bacterial community composition and dynamics spanning five years in freshwater bog lakes', *mSphere*. Am Soc Microbiol, 2(3), pp. e00169--17.

Lipp, e M. and Anklam, E. (1998) 'Review of cocoa butter and alternative fats for use in chocolate-part A. Compositional data', *Food chemistry*. Elsevier, 62(1), pp. 73–97.

Liu, L. *et al.* (2012) 'Comparison of next-generation sequencing systems', *BioMed Research International*. Hindawi Publishing Corporation, 2012.

Liu, L., Zhong, J. and Ma, Y. (2013) 'A multivariate synthetic control chart for monitoring covariance matrix based on conditional entropy', in *The 19th International Conference on Industrial Engineering and Engineering Management*, pp. 99–107.

Lou, Q. and Obradovic, Z. (2012) 'Predicting viral infection by selecting informative biomarkers from temporal high-dimensional gene expression data', in *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*, pp. 1–4.

Love, M. I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and

dispersion for RNA-seq data with DESeq2', *Genome biology*. BioMed Central, 15(12), p. 550.

MacGregor, J. F. and Kourti, T. (1995) 'Statistical process control of multivariate processes', *Control Engineering Practice*. Elsevier, 3(3), pp. 403–414.

Macke, J. H. *et al.* (2009) 'Generating spike trains with specified correlation coefficients', *Neural computation*. MIT Press, 21(2), pp. 397–423.

Mangan, N. M. *et al.* (2016) 'Inferring biological networks by sparse identification of nonlinear dynamics', *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*. IEEE, 2(1), pp. 52–63.

Martín-Fernández, J.-A. *et al.* (2015) 'Bayesian-multiplicative treatment of count zeros in compositional data sets', *Statistical Modelling*. Sage Publications Sage India: New Delhi, India, 15(2), pp. 134–158.

Maxam, A. M. and Gilbert, W. (1977) 'A new method for sequencing DNA', *Proceedings of the National Academy of Sciences*. National Acad Sciences, 74(2), pp. 560–564.

Mayer-Schönberger, V. and Cukier, K. (2013) *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Meinshausen, N., Bühlmann, P. and others (2006) 'High-dimensional graphs and variable selection with the lasso', *The annals of statistics*. Institute of Mathematical Statistics, 34(3), pp. 1436–1462.

Metzker, M. L. (2010) 'Sequencing technologies - the next generation', *Nature reviews genetics*. Nature Publishing Group, 11(1), p. 31.

Miller, P., Swanson, R. E. and Heckler, C. E. (1998) 'Contribution plots: A missing link in multivariate quality control', *Applied mathematics and computer science*, 8(4), pp. 775–792.

Miriello, D. *et al.* (2010) 'Characterisation of archaeological mortars from Pompeii (Campania, Italy) and identification of construction phases by compositional data analysis',

Journal of Archaeological Science. Elsevier, 37(9), pp. 2207–2223.

Modarres, R. and Jernigan, R. W. (1993) 'A robust test for comparing correlation matrices', *Journal of statistical computation and simulation*. Taylor & Francis, 46(3–4), pp. 169–181.

Montagud, A. *et al.* (2019) 'Conceptual and computational framework for logical modelling of biological networks deregulated in diseases', *Briefings in bioinformatics*. Oxford University Press, 20(4), pp. 1238–1249.

Muetze, T. *et al.* (2016) 'Contextual Hub Analysis Tool (CHAT): A Cytoscape app for identifying contextually relevant hubs in biological networks', *F1000Research*. Faculty of 1000 Ltd, 5.

Mullis, K. *et al.* (1986) 'Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction', in *Cold Spring Harbor symposia on quantitative biology*, pp. 263–273.

Munos, B. (2009) 'Lessons from 60 years of pharmaceutical innovation', *Nature reviews Drug discovery*. Nature Publishing Group, 8(12), pp. 959–968.

Nguyen, N.-P. *et al.* (2016) 'A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity', *NPJ biofilms and microbiomes*. Nature Publishing Group, 2, p. 16004.

Paul, S. M. *et al.* (2010) 'How to improve R&D productivity: the pharmaceutical industry's grand challenge', *Nature reviews Drug discovery*. Nature Publishing Group, 9(3), pp. 203–214.

Pearson, K. (1897) 'Mathematical contributions to the theory of evolution.—on a form of spurious correlation which may arise when indices are used in the measurement of organs', *Proceedings of the royal society of london*. The Royal Society, 60(359–367), pp. 489–498.

Pinzón, H. *et al.* (2018) 'A Novel Hybrid Strategy for Multimode Operation Mapping and Feature Extraction on Data-Driven Statistical Fault Detection Methods', in *ASME 2018 International Mechanical Engineering Congress and Exposition*.

Portnoy, I. *et al.* (2016) 'An improved weighted recursive PCA algorithm for adaptive fault detection', *Control Engineering Practice*, 50, pp. 69–83. doi: 10.1016/j.conengprac.2016.02.010.

Qiu, H. *et al.* (2016) 'Joint estimation of multiple graphical models from high dimensional time series', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. Wiley Online Library, 78(2), pp. 487–504.

Rato, T. J. and Reis, M. S. (2014) 'Sensitivity enhancing transformations for monitoring the process correlation structure', *Journal of Process Control*. Elsevier, 24(6), pp. 905–915.

Reineberg, A. E. *et al.* (2018) 'The relationship between resting state network connectivity and individual differences in executive functions', *Frontiers in psychology*. Frontiers, 9, p. 1600.

Reis, M. and Gins, G. (2017) 'Industrial process monitoring in the big data/industry 4.0 era: From detection, to diagnosis, to prognosis', *Processes*. Multidisciplinary Digital Publishing Institute, 5(3), p. 35.

Ren, Z. *et al.* (2015) 'Asymptotic normality and optimalities in estimation of large Gaussian graphical models', *The Annals of Statistics*. Institute of Mathematical Statistics, 43(3), pp. 991–1026.

Roberts, R. J., Carneiro, M. O. and Schatz, M. C. (2013) 'The advantages of SMRT sequencing', *Genome biology*. BioMed Central, 14(6), p. 405.

Roberts, R. J., Carneiro, M. O. and Schatz, M. C. (2017) 'Erratum to: The advantages of SMRT sequencing', *Genome biology*. BioMed Central, 18(1), p. 156.

Rothberg, J. M. *et al.* (2011) 'An integrated semiconductor device enabling non-optical genome sequencing', *Nature*. Nature Publishing Group, 475(7356), p. 348.

Russell, E. L., Chiang, L. H. and Braatz, R. D. (2012) *Data-driven methods for fault detection and diagnosis in chemical processes*. Springer Science & Business Media.

Saegusa, T. and Shojaie, A. (2016) 'Joint estimation of precision matrices in heterogeneous populations', *Electronic journal of statistics*. NIH Public Access, 10(1), p. 1341.

Saha, S. *et al.* (2013) 'Gene expression data clustering using a multiobjective symmetry based clustering technique', *Computers in biology and medicine*. Elsevier, 43(11), pp. 1965–1977.

San-Juan-Vergara, H. *et al.* (2018) 'A Lachnospiraceae-dominated bacterial signature in the fecal microbiota of HIV-infected individuals from Colombia, South America', *Scientific reports*. Nature Publishing Group, 8(1), p. 4479.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977) 'DNA sequencing with chain-terminating inhibitors', *Proceedings of the national academy of sciences*. National Acad Sciences, 74(12), pp. 5463–5467.

Segata, N. *et al.* (2011) 'Metagenomic biomarker discovery and explanation', *Genome biology*. BioMed Central, 12(6), p. R60.

Severson, K., Chaiwatanodom, P. and Braatz, R. D. (2016) 'Perspectives on process monitoring of industrial systems', *Annual Reviews in Control*. Elsevier, 42, pp. 190–200.

Shan, Y. and Deng, G. (2009) 'Kernel PCA regression for missing data estimation in DNA microarray analysis', in *Circuits and Systems, 2009. ISCAS 2009. IEEE International Symposium on*, pp. 1477–1480.

Shendure, J. and Ji, H. (2008) 'Next-generation DNA sequencing', *Nature biotechnology*. Nature Publishing Group, 26(10), p. 1135.

Sonawane, A. R. *et al.* (2019) 'Network medicine in the age of biomedical big data', *Frontiers in Genetics*. Frontiers Media SA, 10.

Städler, N. and Mukherjee, S. (2017) 'Two-sample testing in high dimensions', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. Wiley Online Library, 79(1), pp. 225–246.

Treur, J. (2019) 'The ins and outs of network-oriented modeling: from biological networks and mental networks to social networks and beyond', in *Transactions on Computational Collective Intelligence XXXII*. Springer, pp. 120–139.

Vélez, J. I. and Correa, J. C. (2013) 'Una prueba de independencia completa basada en la FDR', *Comunicaciones en Estadística*, 6(2), pp. 109–120.

Venter, J. C. *et al.* (1998) 'Shotgun sequencing of the human genome'. American Association for the Advancement of Science.

Weng, S., Zhang, C. and Zhang, X. (2004) 'PCA-FA: Applying supervised learning to analyze gene expression data', *Tsinghua Science and Technology*. TUP, 9(4), pp. 428–434.

Wu, N. *et al.* (2019) 'Weighted Fused Pathway Graphical Lasso for Joint Estimation of Multiple Gene Networks', *Frontiers in genetics*. Frontiers, 10, p. 623.

Xia, Y. and Li, L. (2017) 'Hypothesis testing of matrix graph model with application to brain connectivity analysis', *Biometrics*. Wiley Online Library, 73(3), pp. 780–791.

Xia, Y. and Sun, J. (2017) 'Hypothesis testing and statistical analysis of microbiome', *Genes & Diseases*. Elsevier, 4(3), pp. 138–148.

Xu, Y., Yang, Jing-yu and Yang, Jian (2004) 'A reformative kernel Fisher discriminant analysis', *Pattern Recognition*. Elsevier, 37(6), pp. 1299–1302.

Zadeh, L. A. (1999) 'Fuzzy Logic Toolbox For Use With Matlab. The MathWorks Inc'. EconomicModel.

Zhang, Y. *et al.* (2014) 'Opportunities for computational techniques for multi-omics integrated personalized medicine', *Tsinghua Science and Technology*. TUP, 19(6), pp. 545–558.

Zhao, X.-M. and Qin, G. (2013) 'Identifying biomarkers with differential analysis', in *Bioinformatics for Diagnosis, Prognosis and Treatment of Complex Diseases*. Springer, pp.

17–31.

Zuo, Y. *et al.* (2016) ‘INDEED: Integrated differential expression and differential network analysis of omic data for biomarker discovery’, *Methods*. Elsevier, 111, pp. 12–20.

ANEXOS

A.1. Implementación en R de la Técnica propuesta para Evaluación de Similitud de Estructuras de Correlación

```
#User must provide X1 and X2
#Loading Required Libraries and Built-in Functions
library(resample)
library(matlib)
library(MASS)
library(zCompositions)
library(Hotelling)
dimred <- function(lambda, threshold=0.95){
  y <- cumsum(lambda)/length(lambda)
  lambda <- lambda[y <= threshold]
  a <- length(lambda)
  a
}
dimred <- compiler:::cmpfun(dimred)
closure <- function(x,k){
  out <- k*x/(sum(x))
}
statistic <- function(X1, X2){
  S1 <- cor(X1)
  S2 <- cor(X2)
  S1[is.na.data.frame(as.data.frame(S1))] <- 0
  S2[is.na.data.frame(as.data.frame(S2))] <- 0
  lambda1 <- eigen(S1)$values
  a1 <- dimred(lambda1)
  V1 <- eigen(S1)$vectors
  lambda2 <- eigen(S2)$values
  V2 <- eigen(S2)$vectors
  a2 <- dimred(lambda2)
  a <- max(a1,a2)
  phi <- 0
  for(i in 1:a){
    phi <- phi+(max(lambda1[i],lambda2[i]))*(lambda1[i]-lambda2[i])*acos(sum(V1[,i]*V2[,i]))
  }
  phi
}
statistic <- compiler:::cmpfun(statistic)
bS <- function(x1, x2, B = 10000){
  n1 <- NROW(x1)
  n2 <- NROW(x2)
  X <- rbind(x1, x2)
  group <- rep(1:2, c(n1, n2))  # create a variable to identify groups
```

```
  out <- lapply(1:B, function(b){
    status <- sample(group, replace = TRUE)
    x1.samp <- X[status == 1, ]
    x2.samp <- X[status == 2, ]
    statistic(x1.samp, x2.samp)
  })

  unlist(out)
}
bS <- compiler:::cmpfun(bS)

etl <- function(X, isCompositional){
  if(any(X == 0) & isCompositional == TRUE) X <- cmultRepl(X, suppress.print = TRUE)
  if(isCompositional == TRUE) X <- Hotelling::clr(X)
  X
}
etl <- compiler:::cmpfun(etl)

CorTest <- function(X1,X2,n_boot, isCompositional = FALSE, CreateHist = TRUE){
  X1 <- etl(X1, isCompositional)
  X2 <- etl(X2, isCompositional)

  stat <- statistic(X1, X2) ## Calculation of the test proposed statistic
  res <- bS(X1, X2, n_boot)
  if(CreateHist == TRUE){
    hist(res, yaxt = "n", prob = TRUE, ylab = "", breaks = 20, col = 1, border = 1, xlab = 'statistic', main
= '')
    #abline(v = abs(stat), col = 2, lwd = 2)
    abline(v = stat, col = 2, lwd = 2)
  }
  p <- mean(abs(res) >= abs(stat))
  p
}
CorTest <- compiler:::cmpfun(CorTest)
p_value <- CorTest(X1,X2,n_boot=1000, isCompositional=FALSE, CreateHist=TRUE)
```

A.2. Implementación en R del Experimento para Evaluación del Desempeño de las Técnicas para Evaluación de Similitud de Estructuras de Correlación

```
library(psych)
library(PoisNor)
library(zCompositions)
library(Hotelling)
MatGen <- function(n,m,corMat,lamvec){
  no.pois=m
  no.norm=0
```

```
  cmat=corMat
  sd.vec=NULL
  mean.vec=NULL
  x <- PoisNor::genPoisNor(n,no.norm,no.pois,cmat,lamvec,sd.vec,mean.vec)
  x
}
MatGen <- compiler:::cmpfun(MatGen)
dna_test <- function(X1,X2,n_perm){
  min.module.size = 5
  epsilon = 0.5
  num.permutations = n_perm
  rescale.scores = FALSE
  s1 = dna::cornet(X1, rescale.scores)
  s2 = dna::cornet(X2, rescale.scores)
  n1 = nrow(X1)
  n2 = nrow(X2)
  X = rbind(X1, X2)
  p = ncol(X)
  n = n1 + n2
  modF1 = dna::network.modules(abs(s1), min.module.size, epsilon)
  modF2 = dna::network.modules(abs(s2), min.module.size, epsilon)
  F1 = dna::get.modules(modF1)
  F2 = dna::get.modules(modF2)
  G0 = (F1 != 0) | (F2 != 0)
  sNc = 0
  for (g in which(G0)) if ((F1[g] != 0) & (F2[g] != 0)) sNc <- sNc +(F1[g] != 0) *(F2[g] != 0)
*length(intersect(which(F1 ==F1[g]),which(F2 == F2[g])))/length(union(which(F1 ==F1[g]),which(F2
== F2[g])))
  if (sum(G0) == 0)
    sN = 0
  else sN = 1 - sNc/sum(G0)
  names(sN) = NULL
  permutation.list = sample(1:n, n1)
  for (i in 2:num.permutations) permutation.list = rbind(permutation.list,
                                        sample(1:n, n1))
  num.perm = nrow(permutation.list)
  perm.sN = rep(0, num.perm)
  i = 1
  perm.sG0 = rep(0, num.perm)
  perm.mF1 = rep(0, num.perm)
  perm.mF2 = rep(0, num.perm)
  perm.sF1 = rep(0, num.perm)
  perm.sF2 = rep(0, num.perm)
  for (i in 1:num.perm){
    i1 = as.vector(permutation.list[i, ])
    perm.X1 = X[i1, ]
    perm.X2 = X[-i1, ]
    perm.s1 = dna::cornet(perm.X1, rescale.scores)
    perm.s2 = dna::cornet(perm.X2, rescale.scores)
    perm.F1 = dna::get.modules(dna::network.modules(abs(perm.s1), min.module.size,epsilon))
    perm.F2 = dna::get.modules(dna::network.modules(abs(perm.s2), min.module.size,epsilon))
```

```
    perm.G0 = (perm.F1 != 0) | (perm.F2 != 0)
    sNc = 0
    for (g in which(perm.G0)) if ((perm.F1[g] != 0) & (perm.F2[g] !=0)) sNc = sNc + (perm.F1[g] != 0)
* (perm.F2[g] != 0) *length(intersect(which(perm.F1 == perm.F1[g]),which(perm.F2 ==
perm.F2[g])))/length(union(which(perm.F1 ==perm.F1[g]), which(perm.F2 == perm.F2[g])))
    if (sum(perm.G0) == 0) perm.sN[i] = 0
    else perm.sN[i] = 1 - sNc/sum(perm.G0)
  }
  p.value = mean(perm.sN >= sN)
}
dna_test <- compiler:::cmpfun(dna_test)
dimred <- function(lambda, threshold=0.95){
  y <- cumsum(lambda)/length(lambda)
  lambda <- lambda[y <= threshold]
  a <- length(lambda)
  a
}
dimred <- compiler:::cmpfun(dimred)
statistic <- function(x1, x2){
  rho1 <- cor(x1)
  lambda1 <- eigen(rho1)$values
  V1 <- eigen(rho1)$vectors
  a1 <- dimred(lambda1)
  rho2 <- cor(x2)
  lambda2 <- eigen(rho2)$values
  V2 <- eigen(rho2)$vectors
  a2 <- dimred(lambda2)
  a <- max(a1, a2)
  phi <- 0
  for(i in 1:a){
    phi <- phi+(max(lambda1[i],lambda2[i]))*(lambda1[i]-lambda2[i])*acos(sum(V1[,i]*V2[,i]))
  }
  phi
}
statistic <- compiler:::cmpfun(statistic)
statistic_k <- function(x1, x2){
  rho1 <- cor(x1)
  lambda1 <- eigen(rho1)$values
  rho2 <- cor(x2)
  lambda2 <- eigen(rho2)$values
  phi <- sum(abs(lambda2-lambda1))
  phi
}
statistic_k <- compiler:::cmpfun(statistic_k)
bS <- function(x1, x2, n_boot = 10000){
  n1 <- NROW(x1)
  n2 <- NROW(x2)
  X <- rbind(x1, x2)
  group <- rep(1:2, c(n1, n2))  # create a variable to identify groups
  out <- lapply(1:n_boot, function(b){
    status <- sample(group, replace = TRUE)
```

```
    x1.samp <- X[status == 1, ]
    x2.samp <- X[status == 2, ]
    statistic(x1.samp, x2.samp)
  })
  unlist(out)
}
bS <- compiler:::cmpfun(bS)
bS_k <- function(x1, x2, n_boot = 10000){
  # set up
  n1 <- NROW(x1)
  n2 <- NROW(x2)
  X <- rbind(x1, x2)
  group <- rep(1:2, c(n1, n2))  # create a variable to identify groups
  out <- lapply(1:n_boot, function(b){
    status <- sample(group, replace = F)
    x1.samp <- X[status == 1, ]
    x2.samp <- X[status == 2, ]
    statistic_k(x1.samp, x2.samp)
  })
  # output
  unlist(out)
}
bS_k <- compiler:::cmpfun(bS_k)
All_Tests <- function(X1, X2, n_boot = 5000, alpha = 0.05){
  stat <- statistic(X1, X2)
  res <- bS(X1, X2, n_boot = n_boot)
  p_nca <- mean(abs(res) >= abs(stat))
  statk <- statistic_k(X1, X2)
  res <- bS_k(X1, X2, n_boot = n_boot)
  p_k <- mean(abs(res) >= abs(statk))

  p_dna <- dna_test(X1,X2,n_boot)
  out <- c(p_nca, p_k, p_dna)
  out <= alpha
}
All_Tests <- compiler:::cmpfun(All_Tests)
Tests_Assessment <- function(n, m, mean = 100, p = 0, n_boot = 5000){
  ## generate correlation matrix
  if(p == 0) {
    corMat1 <- corMat2 <- diag(m)
  } else{
    corMat1 <- diag(m)
    ip  <- matrix(1, nrow = m, ncol = m)
    corMat2 <- (1-p)*corMat1 + p*ip
  }
  ## data generation
  lamvec <- rep(mean, m)
  X1 <- MatGen(n, m, corMat1, lamvec)
  X2 <- MatGen(n, m, corMat2, lamvec)
  colnames(X1) <- colnames(X2) <- paste0('b', 1:ncol(X1))
  out <- All_Tests(X1, X2, n_boot, alpha = 0.05)
```

```
}
Tests_Assessment <- compiler:::cmpfun(Tests_Assessment)
mean <- 100
q=rev(c(20,40,60,80,100,120,140))
l=c(20,60,100,140,180,220,260)
nl <- length(l)
nq <- length(q)
vl <- c()
vq <- c()
for (i in 1:nl){
  vl <- c(vl,rep(l[i],nq))
}
for (j in 1:nl){
  vq <- c(vq,q)
}
lq <- rbind(vl,vq)
llq <- dim(lq)[2]
p=0.1
n_rep=3
Rejection_Rate_NCA <- Rejection_Rate_dna <- Rejection_Rate_k <- rep(0, llq)
for (k in 1:llq){
  r1_nca <- r1_k <- r1_dna <- rep(0,n_rep)
  for(i in 1:n_rep)
  {
    r1 <- Tests_Assessment(lq[1,k],lq[2,k],mean,p, n_boot = 200)
    r1_nca[i] <- r1[1]
    r1_k[i] <- r1[2]
    r1_dna[i] <- r1[3]
  }
 save(r1,file=paste0("resnewStat_",k,".RData"))
  Rejection_Rate_NCA[k]=100*mean(r1_nca)
  Rejection_Rate_k[k]=100*mean(r1_k)
  Rejection_Rate_dna[k]=100*mean(r1_dna)
}
save(Rejection_Rate_NCA, file = "Rejection_Rate_NCA_rho_0.RData")
save(Rejection_Rate_k, file = "Rejection_Rate_k_rho_0.RData")
save(Rejection_Rate_dna, file = "Rejection_Rate_dna_rho_0.RData")
```

A.3. Implementación en Matlab de la Técnica Propuesta para Análisis Diferencial

La implementación consta de un código principal y una serie de funciones, de creación propia del autor, que son llamadas por éste.

A.3.1. Código Principal

```
load Xv_raw; %User defined
load Xc_raw; %User defined
load Tags_filt_tot; %User defined

[nv,mv]=size(Xv_raw); [nc,mc]=size(Xc_raw);
X_raw=[Xc_raw' Xv_raw']'; [n,m]=size(X_raw);
Z=pretreatment(X_raw);
clear Xc_raw Xv_raw
Xc_raw=Z(1:nc,:); Xv_raw=Z(nc+1:nc+nv,:);
[T2,T2_alpha,Q,Q_alpha,a]=PCA_fcn(Xc_raw,Xv_raw);
cont_T=contribution(Z,(1:nc),a,(nc+1:nc+nv),T2_alpha);
wT=(T2./T2_alpha).*(T2>T2_alpha).*(Q<Q_alpha)+(Q./Q_alpha).*(Q>Q_alpha).*(T2<T2_alpha)+(T
2./T2_alpha).*(Q./Q_alpha).*(T2>T2_alpha).*(Q>Q_alpha);
partial_weighted_cont=zeros(m,nv);
for j=1:m
   partial_weighted_cont(j,:)=cont_T(j,:).*wT;%.*wt; %.*w;
end
partial_weighted_cont=partial_weighted_cont';
weighted_cont=k_closure((sum(partial_weighted_cont))',100);
weighted_cont_appendedTags=[weighted_cont';(1:m)]';
[sortedValues,sortIndex]=sort(weighted_cont,'descend');
weighted_cont_sorted(1:mc,1)=weighted_cont(sortIndex,1);
n_top=10;
Top_10_Variables=(Tags_filt_tot(sortIndex(1:n_top,1)))'
figure(1) %subplot(1,2,1)
%title('a')
barh(weighted_cont_sorted(1:n_top,1)');
set(gca,'YTickLabel',Top_10_Variables) % set(gca,'YTickLabel',sortIndex(1:n_top,1))
set(gca,'YTick',1:n_top)
xlabel('Contribution (%)')
```

A.3.2. Funciones

```
function [Ztreat]=pretreatment(X)
 [n,m]=size(X);
  Z=zeros(n,m);
for j=1:n
  Z(j,:)=clr(closure(BM(X(j,:))));
end
Ztreat=Z;


function [out] = clr(x)
```

```
[n,m]=size(x); D=m;
epsilon=zeros(n,m);
g_mean=g(x);
for i=1:D
   epsilon(i)=log(x(i)/g_mean);
end
out=epsilon;

function [c] = closure (z)
k=100; %k=1;
s=sum(z);
c=(k/s)*z;

function [r]=BM(c)
[n,m]=size(c); D=m; r=zeros(n,m);
if n~=1
  display('wrong vector dimensions')
end
n=sum(c);
x=k_closure(c,n);
t=1/D; s=D; %s=D/2;
Sum=0;
for i=1:m
   if x(i)==0
      r(i)=t*(s/(n+s));
      Sum=Sum+t*(s/(n+s));
   end
end
for i=1:m
  if x(i)~=0
      r(i)=x(i)*(1-Sum);
  end
end

function [g_mean] = g (x)
[n,m]=size(x);
product=1; D=m;
for i=1:D
  product=product*(x(i)^(1/D));
end
g_mean=product;

function [c] = k_closure (z,k)
s=sum(z);
c=(k/s)*z;

function [Tsquare,Threshold,Q,Threshold_Q,a]=PCA_fcn(X_train,X_test)
[n,m]=size(X_train);
[a,P,S,Sig_a,b,Sig,D]=PCA_Training(X_train);
[n_test,m_test]=size(X_test); I=ones(n_test,1);
 %Threshold Calculation
```

```
Fo=finv(0.9,a,n-a);
Tsquare_alfa=((a*(n-1)*(n+1))/(n*(n-a)))*Fo;
Threshold=Tsquare_alfa*ones(1,n_test);
%Scaling Test Dataset
X_test=(X_test-I*b)*(Sig^-1);
%Calculating the T^2 Hotelling's Statistic
Tsquare=zeros(1,n_test); K=P*(Sig_a^-2)*P';
for i=1:n_test
   Tsquare(i)=X_test(i,:)*K*X_test(i,:)';
end
% Calculating Q
Q=zeros(1,n_test);  K2=eye(m)-P*P';
for i=1:n_test
   r=K2*X_test(i,:)';
   Q(i)=(r')*r;
end
% Calculating Threshold for Q
Theta1=0; alpha=0.05;
Theta2=0;
Theta3=0;
for i=a+1:m
   Theta1=Theta1+D(i,i)^1;
   Theta2=Theta2+D(i,i)^2;
   Theta3=Theta3+D(i,i)^3;
end
h0=1-(2*Theta1*Theta3)/(3*Theta2^2);
Ca=norminv(1-alpha,0,1);
Qa=Theta1*(((h0*Ca*(2*Theta2)^0.5)/(Theta1))+1+((Theta2*h0*(h0-1))/(Theta1^2)))^(1/h0);
Threshold_Q=Qa*ones(1,n_test);

function [a,P,S,Sig_a,b,Sig,D]=PCA_Training(X)
[n,m]=size(X); b=mean(X); I=ones(n,1);
Sig=zeros(m,m); vars=var(X);
for i=1:m
      Sig(i,i)=sqrt(vars(i));
end
X=(X-I*b)*(Sig^-1);
S=(1/(n-1))*(X')*X;
[V0,D0]=eig(S);
[V,D]=Reorder(V0,D0);
L=diag(D); a=DimRed(L);
P=V(:,1:a);
Sig0=D^0.5;
Sig_a=zeros(a,a);
for j=1:a
    Sig_a(j,j)=Sig0(j,j);
end

function [V_reord,D_reord]=Reorder(V0,D0)
m0=length(D0(1,:));
d=diag(D0); d0=d;
```

```
D=D0;
V=V0;
for i=1:m0
   for j=1:m0
      if d0(i)<d0(j)
         d(i)=d0(j);
         d(j)=d0(i);
         d0=d;
         D(i,i)=D0(j,j);
         D(j,j)=D0(i,i);
         D0=D;
         V(:,i)=V0(:,j);
         V(:,j)=V0(:,i);
         V0=V;
      end
   end
end
for i=1:m0
   D0(i,i)=D(m0-i+1,m0-i+1);
   V0(:,i)=V(:,m0-i+1);
end
V_reord=V0;
D_reord=D0;


function [a]=DimRed(L)
a=0; sum_tot=sum(L); Sum=0; i=1;
while Sum/sum_tot<0.95
   Sum=Sum+L(i); i=i+1;
   a=a+1;
end

function [cont_T]=contribution(X,train_range,a,test_range,T2_alfa)
X_train=X(train_range,:);
[n,m]=size(X);
b=mean(X_train); I=ones(n,1); Sig=zeros(m,m); vars=var(X_train);
for i=1:m
      Sig(i,i)=sqrt(vars(i));
end
X=(X-I*b)*(Sig^-1); S=(1/(n-1))*(X(train_range,:)')*X(train_range,:);
[V,D]=eig(S); [V,D]=Reorder(V,D);
P=V(:,1:a);
Da=D(1:a,1:a);
v=0;
for k=test_range
   r=0; tk=P'*X(k,:)'; v=v+1;
   for i=1:a
      if ((tk(i)^2)/D(i,i))>(T2_alfa(1))^(1/a)
         r=r+1;
          tk_r(r)=tk(i); s(r)=D(i,i); index(r)=i;
      end
```

```
    end
    for j=1:m
        Cont_ij=0;
        for i=1:r
sum=(tk_r(i)/s(i))*P(j,index(i))*(X(k,j));
            if sum>=0
            Cont_ij=Cont_ij+sum; %muh=0
            end
        end
    cont_T(j,v)=Cont_ij;
    end
end
end
```

A.4. Implementación de Experimento para Evaluación del Desempeño de la Técnica Propuesta para Análisis Diferencial

```
clear all; close all; clc; c=50; err=0.02;
N=[20,60,100,140,180,220,260]; M=[20,40,60,80,100,120,140];
n_rep=2; n_exp=length(M)*length(N); Matrix=zeros(n_exp,2); count=0;
for i=1:length(M)
    for j=1:length(N)
            count=count+1;
            Matrix(count,1)=M(1,i); Matrix(count,2)=N(1,j);
    end
end
ade1=zeros(1,n_exp);
for i=1:n_exp
    m=Matrix(i,1); n=Matrix(i,2); b=100*ones(1,m); Sc=eye(m); Sv=Sc;
    bv=b; Sv(1,1)=c; Sv(2,2)=1; Sv(3,3)=1;
    Sv(1,2)=c; Sv(2,1)=c; Sv(1,3)=c; Sv(3,1)=c; Sv(2,3)=c; Sv(3,2)=c;
    de1=zeros(n_rep,1);
    for j=1:n_rep
        Xc_raw=(sampleCovPoisson(b,Sc,n,err))';
        Xv_raw=(sampleCovPoisson(bv,Sv,n,err))';
        weighted_cont=wc(Xc_raw,Xv_raw);
        de1(j,1)=100*(sum(weighted_cont(1,1)>=weighted_cont(2:m,1)))/(m-1);
    end
ade1(1,i)=mean(de1);
end
Results=[Matrix ade1'];
```